Statistical methods for Clinical Genome Interpretation with specific application to Inherited Cardiac Conditions

Imperial College London

Xiaolei Zhang

This dissertation is submitted for the degree of Doctor of Philosophy in Clinical Medicine Research, Imperial College London.

April 2021

Abstract

Background: While next-generation sequencing has enabled us to rapidly identify sequence variants, clinical application is limited by our ability to determine which rare variants impact disease risk.

Aim: Developing computational methods to identify clinically important variants

Methods and Results:

(1) I built a disease-specific variant classifier for inherited cardiac conditions (ICCs), which outperforms genome-wide tools in a wide range of benchmarking. It discriminates pathogenic variants from benign variants with global accuracy improved by 4-24% over existing tools. Variants classified with >90% confidence are significantly associated with both disease status and clinical outcomes.

(2) To better interpret missense variants, I examined evolutionarily equivalent residues across protein domain families, to identify positions intolerant of variations. Homologous residue constraint is a strong predictor of variant pathogenicity. It can identify a subset of *de novo* missense variants with comparable impact on developmental disorders as protein-truncating variants. Independent from existing approaches, it can also improve the prioritisation of disease-relevant gene for both developmental disorders and inherited hypertrophic cardiomyopathy.

(3) TTN-truncating variants are known to cause dilated cardiomyopathy, but the effect of missense variants is poorly understood. Using the approach in (2), I studied the role of *TTN* missense variants on DCM. Our prioritised residues are enriched with known pathogenic variants, including the two known to cause DCM and others involved in skeletal myopathies. I

also found a significant association between constrained variants of TTN I-set domains and DCM in a case-control burden test of Caucasian samples (OR=3.2, 95%CI=1.3-9.4). Within subsets of DCM, the association is replicated in alcoholic cardiomyopathy.

(4) Finally, I also developed a tool to annotate 5'UTR variants creating or disrupting upstream open reading frames (uORF). Its utility is demonstrated to detect high-impact uORF-disturbing variants from ClinVar, gnomAD and Genomics England.

Conclusion:

These studies established broadly applicable methods and improved understanding of ICCs.

Declaration

Statement of Originality

I hereby certify that the submission is my original work. To the best of my knowledge, it does not contain sources or resources other than the ones that have been appropriately acknowledged or referenced. It has not been submitted previously for any degree.

Copyright Declaration

This thesis is dedicated to my mother Yuyuan Liu and father Zhongyun Zhang, who adopted me, gave me the best love that I can never return back, and completely changed my destiny.

这本博士论文献给我的妈妈刘玉媛和爸爸张中远，因为他们收养了我，给了我永远也无法回报的爱护，彻底地改变了我的人生命运。

Acknowledgements

What makes a PhD?

For my PhD, I wouldn't be able to come to this point without the influence and support from everyone.

Thank you to my first supervisor, James Ware, who always questions critically and has helped me to think beyond the results from single pieces of work and remember the impact we want to make by doing research: “Is it really useful in clinics?”, “What would you tell people on the street if they have this variant?”. Meanwhile, he shows me what is the basis underlying significant work, with his scientific meticulousness: “Are we really using the right statistical test?”, “How do you define rare? why is it 0.1%, or not?”. Thank you for providing me the incredible guidance to work on challenging questions, as well as support and academic freedom to explore my own research interests.

I am also grateful to have Nicky Whiffin as my supervisor. Thank you for your mentorship, which has helped me get back on track several times during my PhD. You remind me that I can be the best of myself. Your strengths and warmth have greatly inspired me that we can be the scientists with our own characters.

Thank you to my supervisor Leonardo Bottolo, for your valuable guidance and support on statistics. Your enthusiasm over tricky questions shows me that doing science can be so cheerful.

I am also thankful for being in a community as supportive as the Genetics and Genomics Lab. Thank you to Paul Barton, who has always opened his office door for us (that was pre-Covid; during Covid, he has made his calendar open), gave us a place and listened to us: “You are not the only one who sit and cry here”. Thank you for helping me through the challenging times.

I am also grateful to be in a community of enthusiastic, brilliant, and caring peers, who have helped and advised me scientifically or personally, or shared their passion for science, especially Risha, Roddy, Alicja, Erica, Mian, Francesco, Mikyung, Rachel, Mona, Ang, Kathryn, Chulin, Xiao, Panda, Nick Li, Nick Quaife, Catherine Enright, Declan, Antonio and Sean.

Finally, great thanks to my family and friends, especially my brother, my sister-in-law, my nieces, who have shared the ups and downs with me through PhD.

Table of Contents

Publications arising from this work

The main study of Chapter 2 has been published alongside a method paper:

1. **Zhang, X**., Walsh, R., Whiffin, N. *et al.* Disease-specific variant pathogenicity prediction significantly improves variant interpretation in inherited cardiac conditions. *Genet Med* (2020). *https://doi.org/10.1038/s41436-020-00972-3*

2. **Zhang X**, Minikel EV, O'Donnell-Luria AH *et al.* ClinVar data parsing. *Wellcome Open Res* 2017, **2**:33 *https://doi.org/10.12688/wellcomeopenres.11640.1*

A study based on methods developed in Chapter 4:

3. Tijsen, A., Ortega, L., Reckman, Y. **Zhang, X.** *et.al* TTN circular RNAs create a backsplice motif essential for SRSF10 splicing. *Circulation* 143.15 (2021): 1502-1512. https://doi.org/10.1161/CIRCULATIONAHA.120.050455

Part of Chapter 5 has been published:

4. **Zhang X**, Wakeling M, Ware J, Whiffin N. Annotating high-impact 5'untranslated region variants with the UTRannotator. Bioinformatics. 2020 Sep 14:btaa783. *https://doi.org/10.1093/bioinformatics/btaa783*

5. Whiffin, N., Karczewski, K.J., **Zhang, X**. *et al.* Characterising the loss-of-function impact of 5’ untranslated region variants in 15,708 individuals. *Nat Commun* **11,** 2523 (2020). *https://doi.org/10.1038/s41467-019-10717-9*

Conference Oral presentations

1. Interpretation and understanding of 5’UTR variants. European Society of Human Genetics, 2021 (invited talk).
2. Annotating high-impact 5’untranslated region variants with the UTRannotator. American Society of Human Genetics, 2020.
3. Identification of TTN missense variants with a role in Dilated Cardiomyopathy. Genomics of Rare Diseases, Wellcome Genome Campus. 2020
4. Interpretation of rare variants with applications in Inherited Cardiac Conditions. Leena Peltonen Summer School of Human Genomics. 2019
5. Disease-specific variant pathogenicity prediction using machine learning methods improves variant interpretation in inherited cardiac conditions. The Francis Crick AI Club. 2017

Awards

Finalist of 2020 Charles J. Epstein Trainee Award for Excellence in Human Genetics Research

Abbreviations

ACM	Alcoholic Cardiomyopathy
ACMG	American College of Medical Genetics
AFR	African population
ASD	Autism spectrum disorders
BrS	Brugada syndrome
CAU	Caucasian population
CCR	Constrained Coding Region
CDS	Protein-coding sequence
Chemo	Chemotherapy-induced Cardiomyopathy
CM	(Inherited) Cardiomyopathy
DCM	(Familial) Dilated Cardiomyopathy
DD	Developmental disorders
DM	Disease Mutation
DNV	*De novo* variants
EAS	East Asian population
ESP	NHLBI Exome Sequence Project
ExAC	Exome Aggregation Consortium
FNR	False Negative Rate
FPR	False Positive Rate
GEL	Genomics England
gnomAD	Genome Aggregation Database
GWAS	Genome-wide Association Study
HCM	Hypertrophic Cardiomyopathy
HGMD	Human Gene Mutation Database
HMERF	Autosomal Dominant Hereditary Myopathy with Early Respiratory Failure

HR	Hazard Ratio
HRC	Homologous Residue Constraint
IAS	Inherited Arrhythmia Syndrome
ICC	Inherited Cardiac Condition
LMM	Laboratory of Molecular Medicine
LoF	Loss-of-function
LQTS	Long QT syndrome
MCC	Matthews Correlation Coefficient
MRI	magnetic resonance imaging
NDD	Neurodevelopmental delay
NPV	Negative Predictive Value
OMGL	Oxford Medical Genetics Laboratory
oORF	Overlapping open reading frame
OR	Odds Ratio
ORF	Open reading frame
PPV	Positive Predictive Value
Pr	Probability of pathogenicity
PR-AUC	Area under the Precision-Recall Curve
PTV	Protein-truncating variants
RBH	Royal Brompton & Harefield Hospitals NHS Trust
RMC	Regional Missense Constraint
ROC-AUC	Area under the Receiver Operating Characteristic Curve
SHaRe	Sarcomeric Human Cardiomyopathy Registry
SNV	Single-nucleotide variant
TNR	True Negative Rate
TPR	True Positive Rate
TTNtv	Titin (TTN) truncating variants

UKBB	UK Biobank
uORF	Upstream open reading frame
UTR	Untranslated region of mRNA
VEP	Ensembl Variant Effect Predictor
VUS	Variant of Uncertain Significance
WGS	Whole-genome sequencing

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Computational approaches to interpret genomes of rare diseases

The genetic basis of human diseases can be typically categorised as: polygenic, monogenic diseases, or in-between. Polygenic diseases, also known as complex traits or common diseases, are caused by variants from many genes with small effects. On the contrary, monogenic diseases, also known as Mendelian diseases or rare diseases, are caused by variants in a single gene with a large effect. Throughout my thesis, I focus on studying large-effect variants causing monogenic diseases including inherited cardiac conditions or other rare disorders.

High-throughput sequencing technologies have greatly improved our ability to discover genetic variants. Typically, the number of genetic variants identified are thousands in whole-exome sequencing of an individual and even millions in whole-genome sequencing (WGS)[1]. While not all of them have impacts on diseases, distinguishing the ones that do have a clinical impact from background ones is the central challenge in human disease genetics.

To interpret variant pathogenicity, current clinical practice follows the American College of Medical Genetics (ACMG) guidelines[2], which is a standard semi-quantitative classification framework to integrate diverse lines of evidence including computational, population, segregation, and functional data. Based on combination of strengths of evidence, the classification of variant pathogenicity includes the following five categories ordered by severity: Pathogenic, Likely pathogenic, Variant of Uncertain Significance, Likely Benign, and Benign. In practice, strong genetic evidence such as segregation and functional data is not always available. Thus, robust computational and population evidence is critical to efficiently prioritise variants that have not been seen before in humans. In this section, I am going to summarize *in silico* tools and computational evidence relevant to evaluate variant pathogenicity.

To be clear, throughout this thesis, I reserve the standard terms “Pathogenic” or “Likely pathogenic” for variants that have been evaluated using the full ACMG/AMP framework. I use “disease-causing”, “likely disease-causing” or “deleterious” to indicate pathogenicity predictions from *in silico* tools including the ones I develop and the benchmarked tools.

1.1.1 Variant annotation

To simplify the identification of variant effect, we can categorise variants by predicting consequence on genes or gene products from genomic positions of variants. This is the very first step in genome interpretation. For example, a missense variant denotes a sequence variant that changes one amino acid in a protein but maintains its length, or a start-lost variant means a variant changes the start codon. To standardise the description on the variant consequence, the Sequence Ontology[3] provides a structured vocabulary commonly used in genomics community. This step is automated in popular computational tools such as the Ensembl Variant Effect Predictor (VEP)[4], snpEff[5], and GATK VariantAnnotator[6], which can quickly relate genome annotation (i.e. location of genes) to label variants on the transcripts they disrupt and their predicted consequence on the transcripts. Variant annotation also gives a general order of severity for variant consequences. For example, predicted loss-of-function variants such as start-lost variants have a higher probability of being more deleterious than missense variants or synonymous variants.

However, simply relying on variant consequence is insufficient to classify variants’ clinical impact since a loss-of-function (LoF) variant does not always cause diseases (e.g. a heterozygous LoF variant in disease recessive genes or in redundant genes)[7]. In dominant hypertrophic cardiomyopathy (HCM), a heterozygous LoF variant in *MYH7* is benign while missense variants could cause disease likely acting through dominant-negative mechanism[8]. A small fraction of synonymous variants could have a severe impact by disrupting splicing[9], transcriptional and translational efficiency due to codon usage bias[10].

Therefore, after variant annotation, we want to assess variant-level evidence since even novel variants in established Mendelian genes can also be prevalent in healthy humans, indicating that they are not necessarily causal. For example, *TTN* missense variants can cause dilated cardiomyopathy however each healthy human has around ~50% chance of carrying a rare *TTN* missense variant (presented with data in Chapter 4). For variants found in genes without an established causal link with disease such as in scenarios of gene discovery, we would want to include both gene-level evidence and variant-level evidence in the assessment.

There have been diverse computational approaches and methods developed to assess variant pathogenicity based on gene-level and variant-level evidence. Some approaches are specific for variants of certain consequences depending on the molecular/cellular mechanisms while the others are generalisable for different consequences. For example, sequence conservation is widely adopted to assess variants from coding to non-coding but predicting the effect on protein structure is specific to assess the consequence of missense variants. In the scope of my thesis, I am going to introduce the main categories of tools to interpret missense variants though most of the principles behind them are also applicable to interpret other genetic variants.

Interpreting missense variants remains a critical challenge in genome medicine research. On one hand, as they only change single amino acids in proteins, many of them are tolerated, with no impact either on protein functions or diseases. On the other hand, some of them are found to cause severe conditions. Most of the missense variants remain uninterpreted. Given there are 5,206,202 missense variants observed in gnomAD (v2), there are only 332,217 of them reported with clinical significance in ClinVar (up to Nov 2020), leaving ~94% missense variants with unknown clinical impact. This is a lower-bound estimate considering that many missense variants compatible with human life haven't been catalogued.

1.1.2 Sequence Conservation

If a variant is deleterious, it is likely to be under negative selection. One way to infer negative selection is to compare homologous sequences, which is a widely adopted strategy to estimate the deleteriousness of variants. To quantifying conservation, tools like SIFT[11] measured sequence composition (more frequent, more conserved) while other tools such as GERP[12] and PhyloP[13] compared the deviation between the number of substitutions to neutral expectation (fewer substitutions, more conserved).

The definition of homologous sequences and "phylogenetic scope" used to measure conservation would affect the predictive performance considerably[14]. Conservation based on homologous sequences with longer evolutionary distance or less shared biology would have higher specificity but with compromised sensitivity. For example, variant effect prediction using sequence conservation by comparing from human to yeast might have higher specificity but lower sensitivity than comparing between human and other primate species.

While the above classical tools measure conservation on a single position independently in multiple sequence alignments, recent advances explore applying deep learning to learn conservation pattern from raw sequence either in a supervised[15] (training to learn the pattern from variants of known effect) or unsupervised manner[16] (training to learn without knowing variant effect), which offers an advantage to consider dependencies among residues. Though most of the existing tools measured ortholog conservation, the utility of using paralog has also been demonstrated such as para_zscore[17].

1.1.3 Genetic constraint within human populations

Since conservation-based approaches may not be able to detect regions or functional variants under selection specific to humans, alternatively we could measure purifying selection within human populations, which is referred to as genetic constraint throughout the thesis. The genetic constraint has been shown as strong evidence to prioritise disease genes and variants. This is only feasible with the growth of large-scale reference population databases such as the NHLBI Exome Sequencing Project (ESP)[18], the Exome Aggregation Consortium (ExAC)[19], the Genome Aggregation Database (gnomAD)[20] , and The Trans-Omics for Precision Medicine (known as TOPMed)[21]. The recent release of gnomAD (v2) collects variants seen in 125,748 exomes and 15,708 genomes of unrelated individuals of diverse ancestries, largely enabling assessment of low-frequency genetic variants.

To infer the degree of purifying selection, different statistical models have been developed. One might start by using allele frequency to infer selection and deleteriousness since a deleterious variant is expected to have a lower allele frequency. Indeed, for a penetrant variant causing a dominant Mendelian disorder, its frequency seen in the general population should not be more than the prevalence of the disease it causes. Thus typically, allele frequency less than 0.1% is used to define rare variants considered in variant intepretation[22]. A precise approach to estimate the maximum credible allele frequency given disease prevalence, genetic heterogeneity, and penetrance is described by Whiffin *et.al*[23]*.* While we could derive an upper bound of allele frequency to define the variants we consider, below the upper bound the scale of allele frequency does not necessarily reflect the degree of negative selection since allele frequency could be biased by differential nucleotide mutation rates along the genome. For example, CpG-methylated sites are ~10-50 times more mutable than unmethylated CpG sites.

To adjust for mutability, we can compare the mutation occurrence with that of neutral variants. Assuming that synonymous variants are predominantly neutral, we can measure dN/dS, the ratio of the rate of nonsynonymous substitutions to the rate of synonymous substitutions, which is a classical approach but might have relatively low statistical power[24]. Common variants could also serve as a neutral control. For example, the Residual Variation Intolerance Score (known as RVIS)[25] used the ESP dataset (6,500 exomes) and assessed the degree of a gene intolerant of variants by evaluating whether the number of common variants observed in a gene is less than expected predicted by the total number of variants observed (reflecting the mutational burden of a gene). Alternatively, we could also estimate a single-base mutability explicitly using a sequence-context model[24], which means the number of neutral single nucleotide variants in a gene could be predicted. Thus, a gene-level constraint could be measured by comparing the number of rare variants observed to the number of rare neutral variants expected. This approach is adopted to develop pLI score using ExAC dataset (60,706 exomes)[19] and later LOEUF score using gnomAD (125,748 exomes)[20].

Apart from measuring gene-level constraint to prioritise disease genes, identifying sub-genic regions under depletion of variants could be helpful to predict where disease-causing variants would locate within genes, which is demonstrated in Regional Missense Constraint (RMC)[26], Constrained Coding Region (CCR)[27], sub-region RVIS[28] and Missense Tolerance Ratio (known as MTR)[29]. With the increase of whole-genome sequencing samples, non-coding regions intolerant of genetic variation could also be prioritised as shown previously in the evaluation of regulatory elements relevant to developmental disorders[30] and high-impact 5'UTR variants[31], and through scanning across the entire genome[32].

These gene-level or sub-genic level constraint scores have shown to be effective to identify haploinsufficient genes and improve the interpretation of protein-truncating variants and missense variants. As these scores aim to measure the selection force, they might not be the most suitable approach to uncover the variants or genes involved in diseases that occur after reproductive age though empirically common variants in constrained genes are found to be correlated with disease risk[20].

1.1.4 Predicting structural effect of variants

If assessing conservation and genetic constraint is an approach to infer deleteriousness from its consequence (reduce reproductive fitness thus experience purifying selection), another independent resource is to predict variant effect from the molecular mechanism. For missense variants, that means estimating how likely a missense variant would disrupt the structure and function of a protein, and interaction with others. For splice-altering variants, that means evaluating whether a variant would disrupt sequence patterns recognised by a spliceosome.

To gain functional insights of missense variants causing diseases, various studies have found disease-associated structural effects including disrupting protein stability (e.g. breaking a disulfide bond, introducing a hydrophilic residue into the hydrophobic core), perturbing the interaction between protein and protein/DNA/RNA, perturbing protein flexibility and modifying the functionally important sites such as binding activities[33–35].

There are important limitations of applying structural effect prediction in variant interpretation. Firstly, a structurally damaging variant is not necessarily disease-causing. “Structurally deleterious” or “functionally deleterious” is not equivalent to “disease-causing”.

Prior knowledge on disease mechanisms would need to be incorporated. For example, we know that loss-of-function missense variants are most likely found in the protein core than the surface and a heterozygous LoF in *MYH7* is benign for hypertrophic cardiomyopathy. Therefore, we could infer that a missense variant disrupting the core of MYH7 may not be predictive of variant pathogenicity. It has also been shown empirically (i.e. the distribution of *MYH7* missense variants on the core and surface is not different either for HCM cases or controls)[36]. Secondly, the effect prediction relies on accurate protein structures, and only ~50% of human protein have structural models including experimentally determined structures and homology-based predicted structures[33,37]. Looking on the bright side, the performance of using predicted structures to assess variants is shown to be comparable with experimental structures[33]. For the remaining half, it is of great interest to develop accurate *ab initio* structural prediction (only using amino-acid sequence without templates) algorithms exemplified by DeepMind AlphaFold[38] to scale up the prediction of structural damaging variants.

1.1.5 Machine-learning based variant pathogenicity prediction

As mentioned above, ACMG guidelines provide a decision framework to classify variant pathogenicity given multiple lines of evidence, which is a consensus of expert opinions. Apart from relying on human experts to curate diagnosis criteria, we could also learn decision rules and patterns automatically from data by using machine learning algorithms. In a nutshell, machine learning has been applied in variant pathogenicity prediction in two different ways.

Firstly, given multiple lines of evidence, machine learning can be used to find their optimal weights and combination relevant to pathogenicity prediction. State-of-the-art machine learning algorithms are developed to fit different underlying distributions of data. Depending on whether the final prediction output (also known as label) is used, there are two main types of machine learning tasks: supervised learning and unsupervised learning.

The goal of supervised learning is to learn the best approximation about the relationship between input features (lines of evidence) and output label data (whether a variant is pathogenic). For example, ACMG guidelines classify variants as likely pathogenic with one line of strong evidence (e.g. well-established functional studies showing a deleterious variant) and 1-2 line(s) of moderate evidence (e.g. absent in control population datasets). With sufficient data, we could infer the strength of each line of evidence quantitively and the best combination of them from known pathogenic and benign variants. In practice, some lines of evidence mentioned in ACMG guidelines are not always available (e.g., functional and segregation data) thus it's infeasible to implement with all relevant experimental and computational evidence currently equivalent with expert classification (explained in Chapter 2 discussion). Instead, multiple computational lines of evidence could be collected such as what have been introduced above: conservation, genetic constraint, and structural information of residues. To find the best combination of these lines of evidence, recent variant pathogenicity tools used state-of the-art machine learning algorithms such as ensemble learning (e.g. M-CAP[39] and REVEL[40]) and deep neural network (e.g. MVP[41]), which generally show improved classification performance over pre-existing tools due to their flexibility to fit non-linear relationship between input and output. Alternatively, unsupervised learning is also applied in pathogenicity prediction such as Eigen[42] and EVE[16]. Since benign and pathogenic variants shall have different distribution over input features (e.g., the distribution of genetic constraint is skewed with opposite directions in pathogenic and benign variants), the goal of unsupervised learning could be understood as learning the hidden structure (approximates variant pathogenicity) that generates the distributions of features seen in data.

Secondly, recent advances in deep learning especially representation learning techniques could be used to learn evidence directly from raw data input such as sequence. This approach helps to automate the process of generating features (line of evidence). It also offers flexibility in modelling interactions within input data. For example, conventionally sequence conservation is measured either heuristically or through explicit modelling of molecular evolution. With large-scale training data, deep learning could directly learn computational lines of evidence from raw sequences as shown in PrimateAI[15], which learned conservation patterns from multiple sequence alignments from variants of known pathogenicity. Different from classical approaches that assume independence across amino acids, PrimateAI also took account of the surrounding sequence (a sliding window of 51 amino acids) of a query amino acid to capture interactions among sites. This approach is more promised if the underlying allelic mechanism primarily depends on raw sequence content. For example, SpliceAI[43] used deep neural network to recognise splicing motifs from pre-mRNA sequence, which is a highly effective method to predict splice-altering variants.

While machine learning algorithms has greatly improved the accuracy of variant pathogenicity predictions, they are not a panacea for all challenges in variant interpretation. In addition to the design of a machine-learning model, the utility of the model also largely depends on the amount and quality of training data. If the training dataset is small or could not represent all possible patterns, the model would have poor performance once applied to unseen data. For example, for interpretation of missense variants, there are multiple possible allelic mechanisms such as gain-of-function or loss-of-function. It would be questionable how well a machine learning model could predict gain-of-function variants if it's predominantly trained on loss-of-function variants.

1.2 Genetic Basis of Inherited Cardiac Conditions

Inherited cardiac conditions (ICCs) are a crucial burden on human health. It is a blanket term that encompasses a variety of rare genetic disorders of the heart. Classically most of the ICC cases are thought to have a monogenic basis although large-scale genome-wide association studies (GWAS) begin to reveal the contribution of common variants. In the thesis, I will study two groups of ICCs as models of Mendelian diseases: familial cardiomyopathies and inherited arrhythmias. Other groups of ICCs are potentially equally important. These genetic conditions are the leading cause of sudden cardiac death in the paediatric and young adult population.

Cardiomyopathies are myocardial disorders, which cause abnormalities of heart tissue both functionally and structurally. More than half of the patients with sudden cardiac death or heart transplantation in the age group < 35-year old have cardiomyopathies[44]. The main types of cardiomyopathies include hypertrophic and dilated cardiomyopathy.

1.2.1 Genetics of Dilated Cardiomyopathy

Dilated cardiomyopathy is characterised by left ventricular dilation and impaired contractility (left ventricular ejection fraction less than 45%[45]) without ischaemia or abnormal loading conditions. Its prevalence is estimated at 1 per 250 people in the general population[46], which indicates that around 28 million people globally could be affected by DCM. The cause of DCM can be either genetic or non-genetic. Non-genetic factors include exposure to toxins (drug, alcohol, and chemotherapy), myocarditis, or pregnancy[47]. Increasing studies show that genetic susceptibility can interact with the non-genetic triggers. Typical clinical outcomes of DCM include cardiovascular death, heart failure, and arrhythmias.

Around 25-30% of total DCM cases are familial (with family members diagnosed with DCM or sudden cardiac death)[45]. Most cases of inherited DCM are considered to have a monogenic basis with an autosomal dominant inheritance pattern, although mitochondrial, autosomal recessive, and X-linked recessive inheritances have also been identified[48]. *De novo* mutations can also cause DCM in young patients[49]. The genetic cause, in terms of causal genes and pathogenic variants, is heterogeneous among different families, which could be mapped to multiple biological pathways. The application of next-generation sequencing in the past decade has led to the rapid identification of candidate genes. In the Human Gene Mutation Database (HGMD)[50], more than 60 genes have been reported to be associated with DCM[50]. The release of large reference population datasets has enabled the re-evaluation of gene-disease relationship[50]. The NIH ClinGen[51] is a central resource to curate clinically relevant genes through expert panels. In the assessment of clinical validity of 49 genes by ClinGen, there are 10 genes (*TTN*, *TTNT2, TNNC1, SCN5A, RBM20, MYH7, LMNA, FLNC, DES, and BAG3*) with definitive evidence, 1 gene (*DSP*) with strong evidence and 6 genes (*ACTC1, JPH2, NEXN, TNNI3, TPM1, and VCL*) with moderate evidence to cause DCM. Genetic variants from these definitive genes account for about 26% of the familial and early-onset DCM, which is shown in a recent analysis of the largest sequenced DCM cohorts so far[50]. In Chapter 2, I am going to use the above 17 genes with at least moderate evidence to define DCM disease genes.

At the variant level, TTN-truncating variants (TTNtv) are the most common genetic cause of DCM. Apart from familial DCM, TTNtv is also the predominant genetic cause of alcoholic cardiomyopathy, chemotherapy-induced cardiomyopathy and peripartum cardiomyopathy[52]. A detailed introduction on the role of *TTN* variants in DCM is described in Chapter 4.

1.2.2 Genetics of Hypertrophic Cardiomyopathy

Hypertrophic cardiomyopathy is characterised by left ventricular hypertrophy with thickened heart muscle in the absence of secondary causes (e.g., pressure overload and hypertension). It is recommended to diagnose adults by the presence of left ventricular end-diastolic wall thickness >= 15mm[53]. It has a prevalence of around 0.2% (1/500) in the general population[54]. Due to age-dependent expression of cardiac hypertrophy, its prevalence is reported to be higher in cohorts with mean age at 60 years old, reaching 0.29%[55].

While the genetic cause of DCM is heterogeneous in terms of underlying biological pathways, HCM is also a heterogeneous condition, but the majority of cases are caused by variants in sarcomere genes (encoding proteins in the sarcomere structure). Sarcomere genes are the most validated genes associated with HCM based on family linkage studies and functional experiments. While there are more than 50 genes ever suggested to cause HCM or used in the genetic testing panel, recent rigours assessment from ClinGen[56] defines 10 genes with definitive evidence including the most established 8 sarcomere genes (*MYBPC3, MYH7, TNNT2, TNNI3, TPM1, ACTC1, MYL2,* and *MYL3*), *PLN* and *FLNC,* one (ALPK3) with strong evidence, 4 (*ACTN2, CSRP3, JPH2,* and *TNNC1*) with moderate evidence. Other 20 genes are defined to have at least moderate evidence causing syndromes involving HCM. In my work of Chapter 2, I only consider the above genes with at least moderate evidence as HCM disease genes.

HCM is associated with a higher risk of death and other adverse clinical outcomes. The Sarcomeric Human Cardiomyopathy (SHaRe) registry is a centralised database collecting genetic and clinical outcome data on patients of HCM from cardiac centres across eight countries. In a recent study of >24,000 patient-years by SHaRe, the mortality was three times higher in patients of HCM than in the general population of similar ages[57]. Patients carrying pathogenic/likely pathogenic variants on sarcomere genes were at 2-fold greater risk for adverse clinical outcomes compared with genotype-negative patients. Patients with variants of uncertain significance on sarcomere genes have intermediate risk[57]. The association between pathogenic/likely pathogenic sarcomere variants and adverse clinical outcomes is also found in the general population shown in a recent analysis on participants of UK Biobank (UKBB)[58].

1.2.3 The role of common variants on cardiomyopathies

After decades of studies on the Mendelian form of cardiomyopathies, only 30%-40% of patients are currently diagnosed with genotype-positive status (with ≥ 1 causative variant) after genetic testing[8,50]. In carriers of rare causative variants, incomplete penetrance and variable expressivity are also observed. Given the above reasons, recent genetic studies have started to investigate the contribution of common variants to the pathogenesis of cardiomyopathies.

For DCM, in a case-control GWAS meta-analysis (currently largest DCM GWAS meta-analysis), Tadros *et.al*[59] found 13 genome-wide significant loci and two shared loci with HCM but with an opposite direction of effect, indicating the shared genetic pathways affecting the risk of cardiomyopathies. Pirruccello *et.al*[60] carried out a GWAS of left ventricular quantitative measurements derived from cardiac magnetic resonance imaging (MRI) in UKBB, which yields greater power to identify 45 novel genetic loci. These GWAS loci are near known Mendelian cardiomyopathy genes. The derived polygenic scores show association with incidence of DCM and variability of cardiac structure and function even among carriers of TTNtv, suggesting the role of common variants on the complex genetic architecture of DCM.

Common variants could also modify the risk of HCM. Two back-to-back case-control genome-wide association studies found 12 and 16 genome-wide significant loci for HCM respectively[59,61]. In the studies, the polygenic background is found to increase the susceptibility of HCM in sarcomere-negative patients (not carrying a rare pathogenic variant in sarcomere genes) but also influence the disease severity in patients carrying pathogenic sarcomere variants.

1.2.4 Genetics of Inherited Arrhythmias

The main types of inherited arrhythmias include Long QT syndrome (LQTS) and Brugada syndrome (BrS). Both syndromes have a prevalence of about 1 in 2000 in the general population[62,63]. Most familial cases are mainly affected by an autosomal-dominant pattern, but not exclusively. Long QT syndrome is caused by ion channel dysfunction in prolonging cellular repolarization. It is typically diagnosed by the presence of a prolonged QT interval on the ECG without secondary causes. While currently there are at least 17 genes reported to cause LQTS, only eight of them have at least moderate evidence for causality in LQTS according to ClinGen[64]. Three genes (*KCNQ1, KCNH2, and SCN5A*) were found to have definitive evidence causing typical LQTS, which account for > 90% genetically diagnosed cases of LQTS. There are another four genes (*CALM1, CALM2, CALM3,* and *TRDN*) with strong or definitive evidence to cause LQTS with atypical features and one gene (*CACNA1C*) with moderate evidence. The current genetic diagnosis yield for LQTS is around 75%[65].

Brugada syndrome is diagnosed by abnormal elevation of ST segment in the right precordial leads accompanied by other criteria (e.g. family history)[63]. While there are more than 20 genes reported to cause Brugada syndrome, a recent assessment of gene validity shows that only *SCN5A* has definitive evidence for causality and all the other genes only have limited or disputed evidence[66]. At variant level, loss-of-function variants in *SCN5A* cause BrS, which accounts for 20% of the familial cases[67].

1.3 Aim of the thesis

Currently, the genetic diagnosis rate is <50% for most rare genetic disorders. As an effort to improve the determination of genetic cause, my thesis focuses on the following three perspectives to develop novel computational tools by mainly using inherited cardiac conditions as my study models.

Firstly, improve the prediction of disease-causing variants using machine learning. While current state-of-the-art *in silico* tools plays a supporting role though not decisive, they are imperfect to apply in genetic diagnosis due to false-positive predictions and false-negative predictions, which could cause conflicts of interpretation and variant unclassified. To address the drawbacks, I hypothesize that a disease-specific predictor performs better than genome-wide tools. To test the hypothesis, I develop a variant classifier to predict how likely a missense variant causes inherited cardiac conditions (Chapter 2).

Secondly, develop novel evidence to interpret missense variants. While the application of machine learning algorithms in variant pathogenicity provides a powerful approach to statistically summarize multiple lines of evidence, novel evidence associated with variant pathogenicity is also needed to improve prediction. Inspired by existing constraint-based methods mining patterns of variants under depletion in human populations on gene or region level, I develop a novel constraint-metric at amino-acid level aiming to find variants under purifying selection that could be missed out by existing approaches (Chapter 3).

As a case study, I particularly apply this novel metric to find *TTN* missense variants with a role in DCM, which is a critical puzzle in cardiovascular genetics (Chapter 4).

Thirdly, improve variant annotation for non-coding variants. Compared with protein-coding variants, interpretation of non-coding variants is much harder due to various reasons: (1) currently annotation of non-coding functional genomic regions is largely undetermined thus affect variant annotation; (2) Different from protein-coding genes, most non-coding elements are not conserved; (3) limited size of interpreted variants hinders the development of machine learning prediction tools.

For coding variants, variant annotation could help us to label variants by their molecular consequence. This could allow us to use the same strategies and evidence to assess variants of the same consequence. However, for non-coding variants, the current annotation is not as informative as for coding variants. As an example to improve non-coding variant annotation, I develop a computational tool to annotate high-impact 5'UTR variants based on their effects on sequence alteration in 5'UTR (Chapter 5)

Chapter 2 Disease-specific variant pathogenicity prediction significantly improves clinical variant interpretation in inherited cardiac conditions

2.1 Introduction

In Chapter 2, I focus on improving variant pathogenicity prediction in inherited cardiac conditions by building a machine learning-based model in a disease-specific framework.

As introduced in Chapter 1, there are various computational tools available to predict variant pathogenicity. They could focus on one line of evidence such as SIFT[11] and GERP[12] that infer deleteriousness from conservation. They could combine multiple lines of evidence including conservation, structural effects, allele frequency info, and pathogenicity scores from existing tools such as M-CAP[39] and REVEL[40], which achieve state-of-the-art classification performance. However, while these tools learn common patterns important to variant effects across the entire genome, their utilities on specific genes or diseases might be compromised. Anderson *et.al* [68] has conducted a systematic analysis to show that the prediction accuracy of the tools differs by disease phenotypes.

I reason there are several disadvantages to applying existing variant pathogenicity tools in clinical variant interpretation. First, a variant pathogenic to one Mendelian disease is likely benign to other Mendelian diseases. Trained on variants pathogenic to any disease, the genome-wide tools show to predict 20%~88% variants benign to inherited cardiac conditions as pathogenic with high predictive probability (P>90%). On the other hand, they do not benefit from specific lines of evidence, which are only available for a subset of well-characterised genes or diseases. It has been previously shown that the addition of such gene and disease-specific evidence into a transparent Bayesian logistic regression framework improves variant interpretation[69]. Moreover, the performances of these computational tools are routinely evaluated using typical classification performance measures which are not necessarily the most relevant in clinical decision making. However, the clinical relevance measures are not well defined and thus poorly assessed in these tools.

Here I hypothesis that disease-specific variant pathogenicity prediction would outperform genome-wide tools by leveraging expert-curated gene and disease-specific data. To examine the hypothesis, inherited cardiac conditions were chosen as examples. I develop a disease-specific variant interpretation tool CardiacBoost to predict the probability of missense variant pathogenic to two inherited cardiac conditions: inherited arrhythmia syndrome (IAS) and familial cardiomyopathies (CM). In this work, I also define high-confidence classification measures of a probabilistic classifier helpful in clinical settings and use these measures to benchmark the classification performances of different tools.

2.2 Methods

2.2.1 Primary training and test data collection

I consider rare missense variants whose allele frequency is less than 0.1%, using gnomAD (v2.0.1) as our reference population. The value at 0.1% is taken as a conservative maximum credible population allele frequency[23] across a range of inherited cardiac conditions, above which variants are unlikely to cause penetrant disease. The predicted molecular consequences of variants were annotated with Ensembl Variant Effect Predictor[4] (version

91.1 for hg19/GRCh37 human genome assembly) on canonical transcripts relevant to heart tissue (**Table 2.1** and **Table 2.2**)

Pathogenic variants in sixteen genes associated with cardiomyopathies (**Table 2.3**) were collected from the targeted sequencing data of 9,007 patients with either HCM or DCM, recruited or referred for diagnostic sequencing at the Royal Brompton & Harefield Hospitals NHS Trust (RBH, UK), Oxford Medical Genetics Laboratories (OMGL, UK)[8], and the Partners Laboratory of Molecular Medicine (LMM, US)[70,71]. The pathogenic variants from RBH and OMGL were interpreted according to ACMG/AMP guidelines. The pathogenic variants from LMM were interpreted using equivalent previously-described clinical-grade variant classification criteria[70,71].

For inherited arrhythmia syndromes, pathogenic variants in seven genes (**Table 2.2**) were extracted from the ClinVar database[72] (ClinVar Full Release 201912), considering only variants with Pathogenic or Likely pathogenic classifications and no conflicting interpretations (Benign or Likely benign).

Rare benign variants for both conditions were collected from the targeted sequencing of 2,090 healthy volunteers. The age range for the healthy volunteer cohort is 5 to 88 years (mean age = 39, SD=15). It included samples recruited from three sites: Royal Brompton Hospital (n=921, range=18-80 years, mean age=39, SD = 13), Egypt Aswan Heart Centre (n=423, range=5-79, mean age = 30, SD=10)[73] and Singapore National Heart Centre (n=746, range=18-88 years, mean age = 45, SD=17). These volunteers were confirmed to have no cardiac history, no family history of, or suggestive of, an inherited cardiac condition, and no evidence of cardiomyopathy or channelopathy on ECG or cardiac MRI. This cohort provides a lower disease prevalence than a general population (i.e. the prevalence of inherited cardiomyopathies and arrhythmias in a general population is estimated at ~0.75% by summing the combined prevalence of HCM, DCM, LQTS, and Brugada syndrome[23]).

Thus, the variants found in their disease panel genes could be considered as highly likely benign for inherited cardiac conditions, while acknowledging the potential for a low background error rate due to incomplete and age-related penetrance.

Three genes are each associated with two related disease phenotypes in the training & test data (*MYH7* and *TNNI3* with hypertrophic and dilated cardiomyopathies; *SCN5A* with two arrhythmia syndromes, LQT & BrS), with distinct variants causing each phenotype. For each of these genes, variants were aggregated so that the model was trained to discriminate disease-causing for either condition versus benign. The phenotype associated with variation in *PLN* does not fit neatly into the clinical definitions of either HCM or DCM[56], so the output of the model for *PLN* variants is interpreted as a probability of variants causing intrinsic cardiomyopathy. For all other genes, the model was exposed to variants associated with just one phenotype (HCM, DCM, BrS, or LQT; see **Table 2.1** and **Table 2.2**).

Table 2.1 Cardiomyopathy-associated genes included in the study.

Gene symbol	Phenotype	Ensembl gene ID	Ensembl transcript ID	Ensembl protein ID
ACTC1	HCM[1]	ENSG00000159251	ENST00000290378	ENSP00000290378
DES	DCM[3] (syndromic)	ENSG00000175084	ENST00000373960	ENSP00000363071
GLA	HCM[3] (syndromic)	ENSG00000102393	ENST00000218516	ENSP00000218516
LAMP2	HCM[3] (syndromic)	ENSG00000005893	ENST00000200639	ENSP00000200639
LMNA	DCM	ENSG00000160789	ENST00000368300	ENSP00000357283
MYBPC3	HCM	ENSG00000134571	ENST00000545968	ENSP00000442795
MYH7	HCM & DCM[1]	ENSG00000092054	ENST00000355349	ENSP00000347507
MYL2	HCM	ENSG00000111245	ENST00000228841	ENSP00000228841
MYL3	HCM	ENSG00000160808	ENST00000395869	ENSP00000379210
PLN	Intrinsic CM[2]	ENSG00000198523	ENST00000357525	ENSP00000350132
PRKAG2	HCM[3] (syndromic)	ENSG00000106617	ENST00000287878	ENSP00000287878
PTPN11	HCM[3] (syndromic)	ENSG00000179295	ENST00000351677	ENSP00000340944
SCN5A	DCM	ENSG00000183873	ENST00000333535	ENSP00000328968
TNNI3	HCM & DCM[1]	ENSG00000129991	ENST00000344887	ENSP00000341838
TNNT2	HCM[1]	ENSG00000118194	ENST00000367318	ENSP00000356287
TPM1	HCM[1]	ENSG00000140416	ENST00000403994	ENSP00000385107

[1] While there are several genes in this table that have been associated with more than one type of cardiomyopathy, e.g. with different variants causing HCM and DCM, our training and

test data included variants associated with just one type of cardiomyopathy for all genes except *MYH7 and TNNI3.* For *MYH7* and *TNNI3*, the output of CardioBoost should be interpreted as “probability of pathogenicity for HCM or DCM”. For other genes associated with more than one subtype, the classifier is trained for a particular disease only, and should be interpreted as such.

[2] The cardiomyopathic phenotype associated with variants in *PLN* does not fit neatly into the clinical definitions of HCM and DCM, so it has been classified under the broader umbrella of intrinsic cardiomyopathy[56].

[3] These conditions typically present with cardiomyopathy in the context of a broader syndromic phenotype, but may also present with isolated heart disease[56].

Table 2.2 Arrhythmia-associated genes included in the study.

Gene symbol	Phenotype	Ensembl gene ID	Ensembl transcript ID	Ensembl protein ID
CACNA1C	Timothy Syndrome (LQT)	ENSG00000151067	ENST00000399655	ENSP00000382563
CALM1	LQT	ENSG00000198668	ENST00000356978	ENSP00000349467
CALM2	LQT	ENSG00000143933	ENST00000272298	ENSP00000272298
CALM3	LQT	ENSG00000160014	ENST00000291295	ENSP00000291295
KCNH2	LQT	ENSG00000055118	ENST00000262186	ENSP00000262186
KCNQ1	LQT	ENSG00000053918	ENST00000155840	ENSP00000155840
SCN5A	LQT & BrS[1]	ENSG00000183873	ENST00000333535	ENSP00000328968

(LQT = Long QT syndrome; BrS = Brugada syndrome)

[1] For *SCN5A*, the output of CardioBoost should be interpreted as “probability of pathogenicity for LQT or BrS”.

Table 2.3 Data sets used for the development of CardioBoost. The number of missense variants in the training and hold-out test datasets is shown for two groups of inherited cardiac conditions.

	Cardiomyopathies			Arrhythmias		
	Pathogenic	Benign	Total	Pathogenic	Benign	Total
Training data set	238	202	440	168	158	326
Test data set	118	100	218	84	79	163
Total	**356**	**302**	**658**	**252**	**237**	**489**

Table 2.4 The training data and hold-out test data grouped by gene used by CardioBoost for cardiomyopathies. The number of missense variants in the training and hold-out test datasets is shown for each gene.

Gene symbol	Training		Test	
	Benign	Pathogenic	Benign	Pathogenic
ACTC1	0	2	1	0
DES	13	3	4	0
GLA	5	5	3	3
LAMP2	5	2	1	0
LMNA	6	10	5	7
MYBPC3	47	19	27	14
MYH7	25	125	13	64
MYL2	1	11	1	1
MYL3	4	3	2	1
PLN	1	2	1	0
PRKAG2	14	2	7	2
PTPN11	8	1	2	1
SCN5A	55	2	27	0
TNNI3	6	23	4	8
TNNT2	8	14	2	8
TPM1	4	14	0	9
Total	**202**	**238**	**100**	**118**

Table 2.5 The training data and hold-out test data grouped by gene used by CardioBoost for arrhythmias. The number of missense variants in the training and hold-out test datasets is shown for each gene.

Gene symbol	Training		Test	
	Benign	Pathogenic	Benign	Pathogenic
CACNA1C	37	4	19	3
CALM1	0	5	0	1
CALM2	0	4	0	6
CALM3	0	3	0	0
KCNH2	33	54	19	22
KCNQ1	12	55	6	31
SCN5A	58	43	26	21
Total	**140**	**168**	**70**	**84**

2.2.2 Additional replication test data collection

To further validate CardioBoost performance on “unseen” data, I collected additional independent data sets which did not overlap with either the training data of CardioBoost, M-CAP, and REVEL or the hold-out test data of CardioBoost.

For cardiomyopathies, these pathogenic test data sets are composed of 129 Pathogenic/Likely Pathogenic variants identified in HCM patients from the SHaRe Registry[57], 15 ClinVar (ClinVar Full Release 201912)[74] variants classified as Pathogenic/Likely Pathogenic for cardiomyopathies with at least two-star review status, and 145 variants of the Disease Mutation (DM) class from HGMD Pro version 201712 after excluding those also seen in HGMD version 2015.2, since these variants were used in the training of M-CAP and REVEL. For arrhythmias, 77 variants reported to be Pathogenic/Likely Pathogenic by OMGL, and 138 variants of the DM class from HGMD Pro version 201712 were collected after excluding those seen in HGMD version 2015.2. For the three calmodulin genes (*CALM1*, *CALM2* and *CALM3*), I also collected variant functional scores from a previous deep mutational scanning study[75]. In this study, a complete functional map for each possible amino acid change in calmodulin protein was generated by employing a high-throughput functional complementation assay in *S.cerevisiae*. Since the three calmodulin genes encode the same protein sequence, the functional map is the same for the three genes. I think this functional map study provides an orthogonal test dataset to validate our prediction because calmodulin protein is highly conserved in eukaryotes. However, I also recognise that the yeast functional assay cannot fully indicate the clinical impact of variants specific to higher organisms[76].

I expect most variants in disease-associated genes identified in gnomAD to be benign for inherited cardiac conditions since the prevalence of inherited cardiomyopathies and arrhythmias in gnomAD should not exceed those in a general population. Since ExAC[19] variants (ExAC version release 0.3, which represents a subset of gnomAD) were used to train M-CAP and REVEL explicitly, we curated a test set of 2,003 gnomAD variants in which the variants seen in ExAC were excluded. Similarly, for arrhythmias, 1,237 gnomAD variants were collected.

2.2.3 Input variant features collection and pre-processing

Feature collection. I combined both variant effect features collected from previous computational tools, and original newly-derived features.

I used ANNOVAR[77] to collect features from published computational tools (**Error! Reference source not found.**). Fourteen conservation or constraint scores of amino acid change were included from BLOSUM62[78], PAM250[78], Grantham Score[79], LRT[80], PhyloP[81], PhastCons[82], SIPHY[83], fitCons[30], GERP++[12], para_zscore[17] and missense badness[26]. To utilise the predictions of existing genome-wide tools, twenty pathogenicity scores were collected from SIFT[11], Polyphen2[84], MutationTaster[85], MutationAssessor[86], FATHMM[87], FATHMM-MKL[87], PROVEAN[88], VEST3[89], CADD[90], DANN[91], MetaSVM[92], MetaLR[92], Eigen[42], M-CAP[39], REVEL[40] and MPC[26].

To incorporate interspecies conservation maximally, I also derived new features measuring evolutionary conservation levels from orthologous sequence alignments of disease genes. Using the multiple alignments of amino acid (AA) sequences of a set of species, for a given missense variant (with the known site, reference AA and alternative AA) four types of features were extracted:

$$Ratio\ of\ Reference\ AA\ = \frac{\text{\#orthologs in the set that have the reference AA at that site}}{\text{\#orthologs in the set that have no gap at that site}}$$

$$Ratio\ of\ Alternative\ AA\ = \frac{\text{\#orthologs in the set that have the alternative AA at that site}}{\text{\#orthologs in the set that have no gap at that site}}$$

$$Ratio\ of\ No-Gap\ = \frac{\text{\#orthologs in the set that have no gap at that site}}{\text{\#orthologs in the set}}$$

$$Ratio\ of\ Orthologs = \frac{\text{#orthologs in the set}}{\text{#species in the set}}$$

I downloaded multiple sequence alignments of orthologous genes from the UCSC hg19 100-way Multiz alignment[93]. The above four scores were calculated for nine different sets of species: (1) all species included in the 100-way alignment; sets of species clade: (2) Primate (3) Euarchontoglires; (4) Laurasiatheria; (5) Afrotheria; (6) Mammal; (7) Aves; (8) Sarcopterygii and (9) Fish (For species in each clade subset see *http://hgdownload.cse.ucsc.edu/goldenpath/hg19/multiz100way/*).

I also derived region-level features from the AA alignment.
$Mean\ Ratio\ of\ Reference\ AA$ measures the average ratio of $Ratio\ of\ Reference\ AA$ among the allele's 10 nearest neighbouring sites. Similarly, $Mean\ Ratio\ of\ No-Gap$ measures the average $Ratio\ of\ No-Gap$ among the allele's 10 nearest neighbouring sites.

Using the alignment of multiple nucleotide sequences, $Ratio\ of\ Reference\ Nucleotide$ and $Ratio\ of\ Alternative\ Nucleotide$ calculate the frequency of reference nucleotide and alternative nucleotide observed in all orthologs given there is no gap at this site respectively. Similarly, $Ratio\ of\ Reference\ Codon$ and $Ratio\ of\ Alternative\ Codon$ are derived as a measure of conservation at the codon level.

Missing features imputation. Variant pathogenicity scores derived from existing genome-wide classifiers and included as features in our model were not available for all variants considered. I estimated these missing values by using condition mean imputation. For test data, missing values were imputed by using the mean derived in the training data[94].

Features normalisation. In total, I collected 76 features per missense variant. After collecting all the features, we conducted a z-score normalisation on the features of the training data.

The features in test data were also standardised using the means and standard variations of the training data.

Table 2.6 Input variant features collected from existing computational tools.

Features	Data type	Description
Grantham score	Integer	Substitution matrix scoring the distance from one amino acid to the other
BLOSUM62	Integer	
PAM250	Integer	
SIFT	Float	Estimate intolerance to variation from closely-related species sequence alignment
Polyphen2	Float x 2	Machine learning method to predict functional effects using structural and sequence features
LRT_score	Float	The original LRT two-sided *P*-value
MutationTaster	Float	Bayes classifier used to predict pathogenicity of variants
MutationAssessor	Float	Predicts functional impact of amino acid substitutions
FATHMM	Float	HMM model to predict functional effects of variants
PROVEAN	Float	Predicts whether an amino acid substitution or indel has an impact on the biological function of a protein
VEST3	Float	Machine learning method to predict variant functional effects
CADD	Float	SVM models to predict pathogenicity for coding and non-coding variants
DANN	Float	Scores whole-genome variants by training a deep neural network
FATHMM-MKL	Float	Machine learning method to predict variant functional effects
MetaSVM	Float	Machine learning method to predict SNVs functional effects
MetaLR	Float	Very similar to MetaSVM, but better interpretable
Eigen	Float x 2	Unsupervised machine learning methods to predict function effects of coding and non-coding variants
M-CAP	Float	Gradient boosting tree to predict functional effects of missense variants
REVEL	Float	Random Forest to predict functional effects of missense variants
GERP++	Float	Identify constrained elements in multiple alignments
PhyloP	Float x 2	Base pair level multi species conservation
Integrated_fitcons	Float	Estimate of fitness consequences
PhastCons	Float x 2	Regional multi species conservation metric
SiPhy	Float	Detect bases under selection based on multiple alignments
paraZscore	Float	Estimate conservation across related proteins within-species from gene paralog
paraZscore_exist	Integer	Indicate whether the paraZscore of a missense variant is available
Missense badness	Float	Measures the increased deleteriousness of amino acid substitutions when they occur in missense-constrained regions
missense badness_exist	Integer	Indicate whether the misbadness score of a variant is available
MPC	Float	Integrated score of misbadness, polyphen-2 and constraint

2.2.4 Defining high-confidence classification performance measures

Existing machine learning variant classification tools adopted a single threshold to discriminate pathogenic and benign variants. However, the choice of this classification threshold is arbitrary and not consistent among different tools, for example M-CAP[39] made a binary classification using a threshold with a 95% true positive rate (see the relevant discussion in: Limitations in applying a high-sensitivity threshold for variant interpretation) and PolyPhen-2[84] made a ternary classification using two thresholds based on false positive rates.

This arbitrary choice of classification threshold might not be optimal in order to control Type I and Type II errors for different applications. Moreover, the use of a high-sensitivity threshold for variant classification is unlikely optimal for clinical interpretation of individual variants. Instead of using classification thresholds derived from a specific classification method/data set, here we adopt high-confidence classification definitions aligned with ACMG/AMP guideline recommendations for clinical practice[2]: the classification of variants into Likely Pathogenic/Pathogenic or Likely Benign/Benign is proposed to be with at least 90% classification certainty. In other words, variants with a pathogenicity score equal to or larger than 0.9 would be classified as “disease-causing” and those with pathogenicity score equal to or smaller than 0.1 are classified as “benign”. Variants with a pathogenicity scores between 0.1 and 0.9 receive an indeterminate classification (variants of unknown significance) (**Figure 2.1**).

With the defined high-certainty classification thresholds, I derive the corresponding confusion matrix **(Figure 2.1**) from which a series of measures of direct clinical relevance can be computed. I use TPR, the proportion of actual pathogenic variants predicted to be disease-causing, and PPV, the proportion of predicted disease-causing variants that are correctly classified, to evaluate the classifier’s ability to classify pathogenic variants. TNR, the proportion of actual benign variants predicted to be benign and NPV, the proportion of predicted benign variants that are correctly classified are used to assess benign classifications correspondingly. Taking both cases together, the accuracy of high-confidence classifications measures the probability that classification in the actionable range is correct. The proportion of clinically indeterminate classifications measures the probability of a variant not classified with clinical confidence. Formulae for each measure of clinical relevance I used are described in the below session.

2.2.5 Limitations in applying a high-sensitivity threshold for variant interpretation

In M-CAP, the authors defined a single low pathogenicity threshold as clinically relevant to predict disease-causing variants such that M-CAP could have a 95% expected true positive rate (sensitivity). Given a data set, while using a low single classification threshold to increase TPR will decrease the number of false negative predictions, the binary classifier would tend to increase the number of false positive predictions (i.e., truly benign variants predicted to be disease-causing) as well. An ideal classification threshold would be the one that minimizes the total sum of the cost of both errors. While one might prioritise sensitivity for variant prioritisation in some contexts, in the context of clinical variant interpretation, we suggest that the cost of a false positive prediction is at least equivalent to, and in most situations higher than, the cost of a false negative prediction. In neglecting to control the Type II error to have a high true positive rate, there would be two negative consequences: (i) Low positive predictive value: this could be demonstrated as the negative correlation between the true positive rate and positive predictive value using the Precision-Recall Curve (**Figure 2.2**a and **Figure 2.2**c); (ii) High false positive rate: this is demonstrated as the positive correlation between the true positive rate and false positive rate (i.e., 1-TNR) (**Figure 2.2**b and **Figure 2.2**d). Even though the ACMG guidelines recommend not to use one computational tool as a sole line of evidence, but to consider the concordance of multiple computational tools for variant interpretation, the application of a computational tool of high TPR but low TNR or high FPR along with other computational tools would still make the clinical interpretation process rather difficult. For example, the disease-causing prediction of a computation tool for a truly benign variant is very likely to conflict with the correct prediction from the other computational tools or the other lines of evidence of pathogenicity. The contradictory evidence would increase the likelihood that the variant is classified as a variant of uncertain significance (VUS).

2.2.6 Calculation of high-confidence classification measures

	Predicted disease-causing	**Predicted benign**	**Indeterminate**	
Actual pathogenic	TP	FN		T
Actual benign	FP	TN		F
	P	N		

Based on the confusion matrix shown above, I calculated the following ratios of clinical relevance in variant interpretation given *n* test variants

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{T}}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{F}}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{F}}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{P}}$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{N}}$$

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{T}}$$

$$\text{Number of high} - \text{confidence classifications} = \mathrm{P} + \mathrm{N}$$

$$\text{Number of indeterminate classifications} = n - (\mathrm{P} + \mathrm{N})$$

$$\text{Proportion of high} - \text{confidence classifications} = \frac{\mathrm{P} + \mathrm{N}}{n}$$

$$\text{Accuracy of high} - \text{confidence classifications} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}$$

$$\text{Overall accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{n}$$

$$\text{Proportion of indeterminate classifications} = \frac{n - (\mathrm{P} + \mathrm{N})}{n},$$

where T: Actual pathogenic, F: Actual benign, P: Predicted disease-causing (Pathogenicity Pr ≥ 0.9), N: Predicted benign (Pr ≤ 0.1), Indeterminate: $0.1 < \mathrm{Pr} < 0.9$, TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative, T = TP + FN, F = FP + TN, P = TP + FP and N = FN + TN.

2.2.7 Machine learning model training and selection

The analyses were conducted using the R environment[23] and the package mlr[24]. I trained and tested representatives of each of the major classes of statistical and machine learning methods in order to obtain the best classification performances over our training data. Neural network methods were not included due to the limited scope for interrogation and interpretation of feature weightings. Classification algorithms included in the analysis are: Classification and Regression Tree (CART)[97], K nearest neighbours (KNN)[98], Elastic Net Logistic Regression (GLMNET)[99], Support Vector Machine with Radial Basis Kernel Function (SVM-RBF)[100], Random Forest (RF)[101], Bayesian Additive Regression Trees (BART)[102], Adaptive Boosting (AdaBoost)[100], Gradient Boosting Tree (GBM)[100] and Extreme Gradient Boosting (XGBoost)[103].

To fine-tune hyperparameters for each model and identify the model with the best generalisation performance (i.e. best prediction performance on “unseen” data), I applied a nested cross-validation[104]. In this nested cross-validation, the inner-test set (also called “validation set” or “development set”) is used to choose the optimal set of hyperparameters for a given classification algorithm. After the classification algorithm is fitted on the inner loop data set, the outer test set is used to select the best-tuned classification algorithm with respect to its performance on “unseen” test data. I used 5-fold cross-validation in the inner cross-validation loop and 10-fold in the outer cross-validation loop.

The selection of the best classification algorithm is not trivial. To this end, I pre-specified the following optimisation goals:

Goal 1: The optimal classifier outperforms genome-wide machine learning variant classification tools on overall classification measured using PR-AUC.

I consider the PR-AUC as a conventional threshold-independent performance measure. In the training process, PR-AUC is chosen as the objective measure in the inner loop for hyperparameter tuning, i.e. for each candidate classification algorithm considered, the hyperparameters that yield the highest PR-AUC are selected. Then the classification performance of each optimised algorithm is assessed using the outer CV loop.

Goal 2: The optimal classifier has the best Matthews Correlation Coefficient[105] (MCC) using the defined 90% high-confidence classification threshold.

Our aim is to find the optimal classifier that balances both Type I and Type II errors at the 90% high-confidence classification thresholds. When I apply the defined high-confidence classification above, variants are classified into one of three categories: disease-causing, benign, and indeterminate. Since the most common application of a genetic diagnosis in cardiogenetic practice is familial evaluation and predictive testing, where management of negative and inconclusive genetic test results are equivalent[106], I group these variants together for the purposes of model selection, and focus on performance at the higher actionable threshold, comparing disease-causing versus non-actionable indeterminate/benign/likely benign.

I use the MCC, a measure of the correlation between observed and predicted binary classifications, which is relatively robust in an imbalanced data set[107], defined as:

$$MCC = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}.$$

A higher MCC reflects a stronger correlation between observed and predicted binary classification, indicative of performance at the ≥ 0.9 thresholds most relevant in this context. Ideally, I would like to select a classifier that performs best on both goals. If there is more than one classifier satisfying both Goal 1 and Goal 2, I pre-specify the selection of the models using Goal 2, given the most immediate relevance to this task.

The performance of each candidate machine learning algorithm and the representative benchmarking genome-wide variant classification tools (M-CAP and REVEL) in the nested cross-validation are shown in **Table 2.7** and **Table 2.8**. For cardiomyopathy variants, as shown in **Table 2.7** the candidate algorithms that outperform M-CAP and REVEL on all standard classification measures to meet Goal 1 were GLMNET, CART, RF, BART, XGBoost, GBM, AdaBoost, KNN and SVM. Since AdaBoost had the highest MCC score to meet Goal 2, it was selected as the best model. Next the best hyperparameter set for AdaBoost ("loss=exponential" and "nu=0.207") was selected using 5-fold cross validation on the whole cardiomyopathy variant training set. The selected model was trained on the whole training set to generate predictions on unseen data.

Similarly, for inherited arrhythmia syndrome variants, AdaBoost was selected as the best-performing candidate (hyperparameters "loss=exponential" and "nu=0.435"). The prediction model was then trained using the whole arrhythmia training set.

Table 2.7 Cross-validated out-of-sample performance for cardiomyopathy variant pathogenicity prediction. I compared nine classification algorithms including best-in-class representatives of all of the major families of machine learning algorithms. AdaBoost was selected with the best cross-validated out-of-sample performance. PR-AUC: Area under the Precision Recall Curve; ROC-AUC: Area under the Receiver Operating Curve; MCC: Matthew Correlation Coefficient.

Method category	Algorithm	PR-AUC (%)	ROC-AUC (%)	Brier score	MCC
Regression	GLMNET	90	88	0.15	0.10
Tree-based	CART	83	81	0.18	0.43
	RF	90	89	0.14	0.36
	BART	91	89	0.14	0.38
Boosting-based	XGBoost	90	87	0.15	0.51
	GBM	87	87	0.15	0.43
	Adaboost	**90**	**88**	**0.14**	**0.58**
Other classification algorithms	KNN	89	88	0.15	0.43
	SVM-RBF	89	87	0.14	0.36
Existing genome-wide classification tools	M-CAP	80	79	0.19	0.35
	REVEL	79	81	0.19	0.25

Table 2.8 Cross-validated out-of-sample performances for arrhythmia variant pathogenicity prediction. I compared nine classification algorithms including best-in-class representatives of the major families of machine learning algorithms. AdaBoost was selected with the best cross-validated out-of-sample performance.

Method category	Algorithm	PR-AUC (%)	ROC-AUC (%)	Brier score	MCC
Regression	GLMNET	91	91	0.12	0.22
Tree-based	CART	82	86	0.14	0.56
	RF	93	92	0.10	0.45
	BART	93	92	0.11	0.43
Boosting-based	XGBoost	88	90	0.12	0.56
	GBM	87	89	0.12	0.60
	Adaboost	**90**	**90**	**0.13**	**0.65**
Other classification algorithm	KNN	92	91	0.12	0.45
	SVM-RBF	92	92	0.10	0.47
Existing genome-wide classification tools	M-CAP	81	85	0.16	0.38
	REVEL	89	90	0.17	0.59

2.2.8 Permutation test to evaluate the significance of a classification measure

Given a performance measure, I used one-sided permutation test[108] to test whether an observed performance measure of one classifier was significantly better than that of the other classifier. The null hypothesis is that the two classifiers perform the same on this measure. The null distribution is estimated by randomly exchanging observations between the classifiers 10,000 times. Here, an observation represents a variant pathogenic probability predicted by a classifier. *P*-value is estimated as the number of times the permuted difference is larger than the observed difference.

2.2.9 Replication without reliance on gold-standard

To ensure robustness to misclassification in the "gold-standard" out-of-sample test data, I employed two orthogonal approaches to assess CardioBoost's discrimination of pathogenic variants and benign variants. First, I compared the proportion of rare variants in individuals with and without disease, and stratified these variants using CardioBoost. I derived the odds ratio (OR), which provides an estimate of gene-disease association.

Second, I compared the survival outcomes of individuals with HCM, stratified by genotypes classified by CardioBoost. I applied CardioBoost to variants found in a cohort of 803 patients with HCM and a rare missense variant in one of eight HCM-associated genes and compared survival with 1,927 genotype-negative HCM patients. I did not consider individuals carrying variants seen in our training data set. The "event-free survival" time (i.e. time until the first major adverse clinical event) was analysed using Kaplan-Meier survival analysis and the Cox hazard-regression model.

2.2.10 Survival analysis

I collected genotype and clinical outcome data for patients with cardiomyopathy from the SHaRe HCM registry (data release 2019Q3).

I included patients with a diagnosis of HCM, at least 1 clinic visit and at least 1 assessment of left ventricular wall thickness, and only one missense variant in any of eight genes encoding sarcomere proteins (*MYBPC3*, *MYH7*, *TNNT2*, *TNNI3*, *TPM1*, *MYL2*, *MYL3*, and *ACTC1*). Variants identified in SHaRe were classified by SHaRe experts according to ACMG/AMP guidelines. Patients with potentially pathogenic variants in genes encoding non-sarcomere proteins were excluded.

The primary outcome measure was a composite comprising: the first occurrence of sudden cardiac death, resuscitated cardiac arrest, appropriate implantable cardioverter-defibrillator therapy, cardiac transplantation, left ventricular assist device implantation, New York Heart Association class III-IV symptoms, all-cause mortality, atrial fibrillation, stroke, or death, as previously described[57].

Patients were censored either at the date of the first event, or at the last follow-up clinical visit if event-free.

2.2.11 The superior performance of CardioBoost is not due to data leakage

Since CardioBoost training and test data may contain variants used as training data for published genome-wide classification tools whose pathogenicity scores were used as input features by CardioBoost, I also assessed whether using indirectly “seen" data would make CardioBoost overfit and outcompete existing genome-wide classifiers. In particular, I considered previously “seen” variants used in training M-CAP and REVEL. M-CAP was trained on variants of Disease Mutation (DM) Class from HGMD version 2015.2 and ExAC. REVEL was trained on variants of DMs from HGMD version 2015.2 and the Exome Sequencing Project (ESP), the Atherosclerosis Risk in Communities (ARIC) study and the 1000 Genomes Project (KGP) (ESP and KGP are contributing projects in ExAC). I extracted a set of "seen" variants from CardioBoost training data if they are ever seen in the DM Class of HGMD version 2015.2 and ExAC. The remaining variants in the training data constitute the set of purely "unseen" data.

I investigated the impact of using "seen" data from two different viewpoints. One is whether "seen" variants have the same classification as those in our training data. In Cardiomyopathies 323 out of 440 training variants were seen before in HGMD or ExAC. For the DM variants reported in HGMD before, 53 out of 206 cases have an opposite classification as in our training data. In Arrhythmias, there are 253 out of 308 variants ever seen in HGMD or ExAC. Among the 170 DM variants reported in HGMD previously, 38 of them have an opposite classification in our training data. This suggests that even if some variants were used in building previously genome-wide classifiers, their classifications are not necessarily correct and thus it makes the prediction tools less accurate.

The second aspect is to assess whether our machine learning tool could still outcompete M-CAP and REVEL on completely "unseen" data. I compared the prediction performance of stratified hold-out test sets: purely "unseen" data and "seen" data (see **Table 2.10** and **Table 2.11**) with the unstratified hold-out test set. The accuracy was used as an overall measure to compare the performance of each dataset. For cardiomyopathies and arrhythmias, the performances of the three datasets were comparable and not significantly different.

Overall, I found out the variants used in previous genome-wide tools were not necessarily accurately classified. The machine learning tool CardioBoost did improve on cardiomyopathy- and arrhythmia-specific prediction both on “seen” and “unseen data” by leveraging over multiple diverse computational pieces of evidence.

2.2.12 Calibration of PPV and NPV

Given a new dataset or testing context, we could estimate the PPV and NPV of a classifier given the proportion of pathogenic variants amongst variants undergoing classification (Variant Proportion):

$$\text{Variant Proportion} = \frac{\text{Number of pathogenic variants}}{\text{Number of pathogenic variants} + \text{Number of benign variants}}$$

$$\text{PPV} = \frac{\text{TPR} \times \text{Variant Proportion}}{\text{TPR} \times \text{Variant Proportion} + \text{FPR} \times (1 - \text{Variant Proportion})}$$

$$\text{NPV} = \frac{\text{TNR} \times (1 - \text{Variant Proportion})}{\text{TNR} \times (1 - \text{Variant Proportion}) + \text{FNR} \times \text{Variant Proportion}}$$

where TPR: True Positive Rate and TNR: True Negative Rate as defined in (1) and (2), respectively.

2.2.13 Estimating the proportion of pathogenic missense variants in a diagnostic series and a general population

In order to estimate the PPV and NPV when applying CardioBoost in a diagnostic series and a general population, we can first estimate the proportion of pathogenic missense variants of these two populations.

Since in variant interpretation, the limitation of false positive prediction is prioritised. Here I want to derive a reasonably conservative estimate of PPVs by assuming that pathogenic missense variants are penetrant and that the burden of rare missense variants in controls provides an estimate of the burden of rare benign missense variants in any population either cases or control. These assumptions would provide the lower bound of the proportion of pathogenic variants, which is the lower bound of PPV based on.

Based on the above assumptions, the proportion of rare pathogenic missense variants, for a given gene or a gene set, amongst variants identified in a group of patients with disorders could be approximated as:

$$\text{Variant proportion in a case series} = \frac{\text{Burden of pathogenic variants in cases}}{\text{Burden of rare variants in cases}}$$

$$\begin{aligned}&\text{Burden of pathogenic variants in cases}\\&\qquad = \text{ Burden of rare variants in cases} - \text{Burden of rare variants in control}\end{aligned}$$

Similarly, the proportion of rare pathogenic missense variants in a general population could be approximated as:

$$\begin{aligned}&\text{Variant proportion in a general population}\\&\qquad = \frac{\text{Burden of pathogenic variants in a general population}}{\text{Burden of rare variants in a general population}}\end{aligned}$$

$$\begin{aligned}&\text{Burden of rare variants in a general population}\\&\qquad = \text{Burden of pathogenic variants in a general population}\\&\qquad + \text{Burden of benign variants in a general population}\end{aligned}$$

$$\text{Burden of pathogenic variants in a general population} =$$
$$\text{Prevalence of disease } \times \text{Burden of pathogenic variants in cases} =$$
$$\text{Prevalence of disease } \times (\text{Burden of rare variants in cases} - \text{Burden of rare variants in control})$$

$$\text{Burden of benign variants in a general population} = \text{Burden of rare variants in control}$$

For cardiomyopathies, here I consider both dilated cardiomyopathy (DCM) and hypertrophic cardiomyopathy (HCM). The disease prevalence for DCM is estimated as 1/250 and 1/500 for HCM[46]. Thus, adding the prevalence of two conditions, the disease prevalence for cardiomyopathies is

$$\frac{1}{250} + \frac{1}{500} \approx 0.006$$

Using cohort studies from OMGL and LMM[8], the burden of rare missense variants in cases is estimated at 27%. *PTPN11* (it was not sequenced in these cohorts and its contribution to cases is assumed to be marginal) was excluded in the analysis here. Using gnomAD[19] reference population as control, the burden of rare missense variants in control was estimated to be 11% by adding the allele frequencies of rare missense variants seen in gnomAD for all cardiomyopathies-related genes (excluding *PTPN11*).

Thus, the proportion of rare missense variants pathogenic to cardiomyopathies in a diagnostic series is estimated at ~ 60%. The proportion of rare missense variants pathogenic to cardiomyopathies in a general population is estimated as ~1%.

Likewise, the proportions of rare missense variants pathogenic to arrhythmias in a diagnostic series and in a general population are estimated as ~71% and ~0.4% respectively. The disease prevalence of arrhythmias in a general population is ~0.2% by adding the disease prevalence of Long QT syndrome (1/2000) and Brugada syndrome (1/1000). Since the arrhythmias-related genes are not widely assessed in large LQTS and Brugada cohort studies[109,110], here I could only consider four arrhythmias-associated genes *KCNE1, KCNH2*, *KCNQ1* and *SCN5A* here from the LQTS and Brugada cohort studies[109,110], which provides us a lower bound of exact variant proportion. The burden of rare missense variants in arrhythmias is estimated at 18%. From the gnomAD database, I estimate the burden of rare missense variants in control (only including *KCNE1, KCNH2*, *KCNQ1* and *SCN5A*) as 5%.

2.3 Results

2.3.1 Building CardioBoost

The data flow diagram from data collection, machine learning model training and testing is illustrated in **Figure 2.1**.

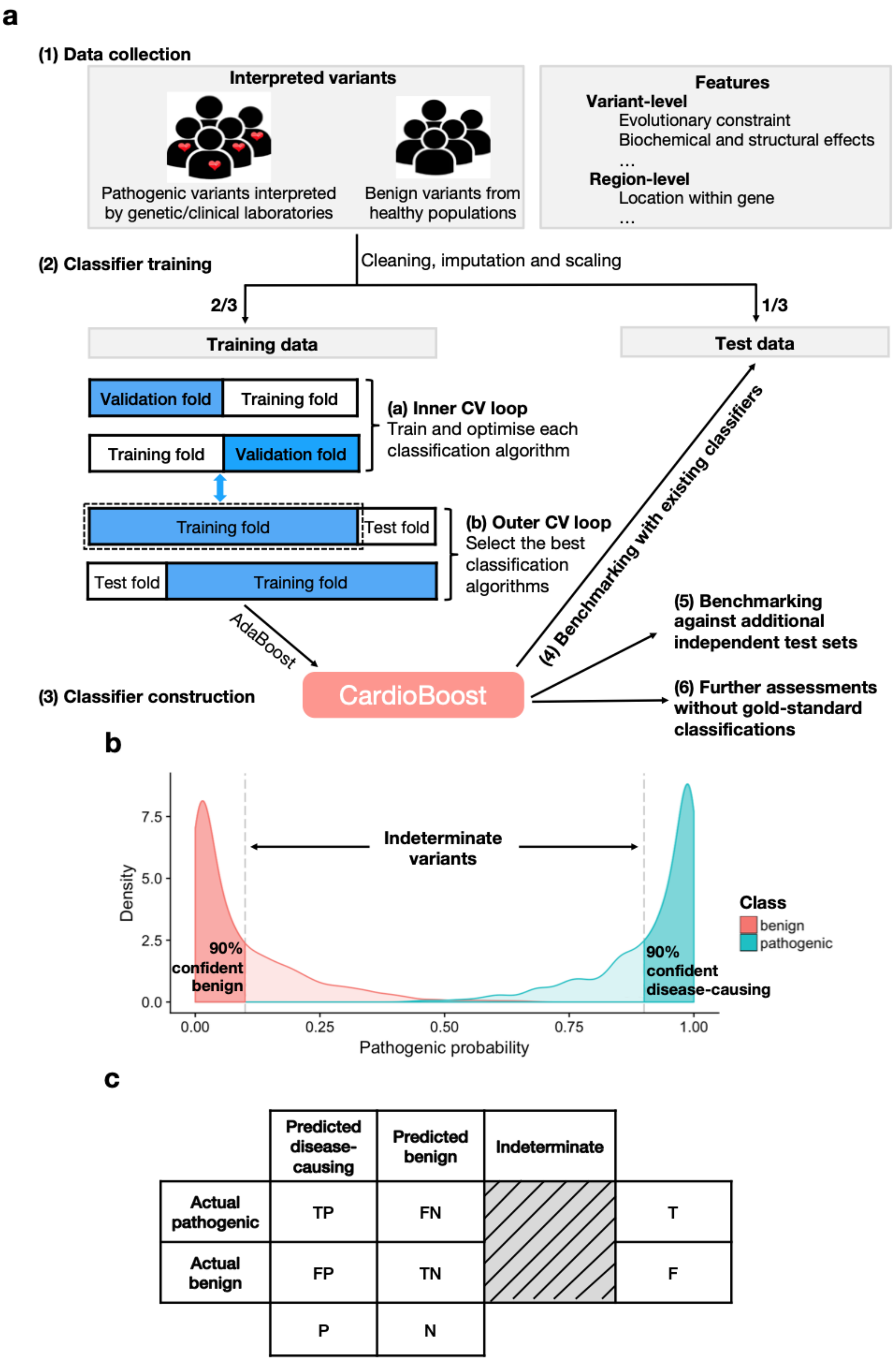
In brief, I constructed two classifiers, one for inherited cardiomyopathies, and one for inherited arrhythmia syndromes, to output the estimated probability of pathogenicity for rare missense variants in genes robustly associated with these conditions.

The CM classifier is applicable for 16 genes associated with hypertrophic and dilated cardiomyopathies. To obtain training and test sets, 356 unique rare (gnomAD minor allele frequency < 0.1%) missense variants were collected in established cardiomyopathy-associated genes (**Table 2.1**) identified in 9,007 individuals with a clinical diagnosis of CM and interpreted as Pathogenic or Likely Pathogenic. For the inherited arrhythmia classifier, I consider genes associated with long QT syndrome and Brugada syndrome. To maximise the size and diversity of the training data, I used ClinVar and only included variants with no conflicting interpretation (Conflicting: P/LP vs B/LB; P/LP vs VUS; B/LB vs VUS). 252 unique rare missense variants reported to be Pathogenic or Likely Pathogenic with no conflicting interpretations (Benign or Likely benign) in established arrhythmia-associated genes (**Table 2.2**) were collected from ClinVar Database[74]. As a benign variant set, 302 unique rare missense variants in cardiomyopathy genes, and 237 unique rare missense variants in arrhythmia genes were collected from the targeted sequencing of 2,090 healthy volunteers. Since these volunteers have no family history of ICCs and are confirmed without ICCs on ECG or cardiac MRI, this cohort provides a lower disease prevalence than a general population thus the rare missense variants carried by them shall be considered as highly likely benign to inherited cardiac conditions. To avoid over-fitting, for each condition the data set was randomly split, with two-thirds used for training and one-third reserved as a hold-out test set (**Table 2.3**, **Table 2.4** and **Table 2.5**). For cardiomyopathies, 440 and 326 variants are used for training and testing respectively. For arrhythmias, 218 and 166 variants are used for training and testing respectively.

For each variant, I collected 76 functional annotations (**Table 2.6**) as features in our disease-specific variant classification tool. I selected nine classification algorithms including best-in-class representatives of all of the major families of machine learning algorithms, and applied a nested cross-validation[104] to select the optimal algorithm for our tool. In the inner 5-fold cross-validation loop, a candidate classification algorithm was trained in order to optimise its hyper-parameters. In the outer 10-fold cross-validation loop, the optimised candidate algorithms were compared and the best-performing one was selected.

For both conditions, AdaBoost[100] was selected with the best cross-validated out-of-sample performance (see **Table 2.7** and **Table 2.8**). AdaBoost is a boosting tree classification algorithm combining many decision trees. Each decision tree is learned sequentially to assign more weight to samples misclassified by the previous decision tree, and weighted by its accuracy. Having selected AdaBoost as the basis for our classifier, a predictive model was constructed by training AdaBoost on the whole set of training variants for each disease, named CardioBoost.

Disease-specific Variant Pathogenicity Classification


a
(1) Data collection
Interpreted variants
Pathogenic variants interpreted by genetic/clinical laboratories
Benign variants from healthy populations
Features
Variant-level
Evolutionary constraint
Biochemical and structural effects
…
Region-level
Location within gene
…
(2) Classifier training
Cleaning, imputation and scaling
2/3
1/3
Training data
Test data
Validation fold
Training fold
Training fold
Validation fold
(a) Inner CV loop
Train and optimise each classification algorithm
Training fold
Test fold
Test fold
Training fold
(b) Outer CV loop
Select the best classification algorithms
AdaBoost
(4) Benchmarking with existing classifiers
(5) Benchmarking against additional independent test sets
(3) Classifier construction
CardioBoost
(6) Further assessments without gold-standard classifications
b
Indeterminate variants
7.5
5.0
2.5
0.0
Density
90% confident benign
90% confident disease-causing
0.00
0.25
0.50
0.75
1.00
Pathogenic probability
Class
benign
pathogenic
c

	Predicted disease-causing	Predicted benign	Indeterminate	
Actual pathogenic	TP	FN	▨	T
Actual benign	FP	TN	▨	F
	P	N		

Figure 2.1 Overview of methods building CardioBoost: training, and testing of CardioBoost, and definition of high-confidence variant classification thresholds for performance assessment. (a) Construction of CardioBoost: (1) After defining gold-standard data, (2) the dataset was split with a 2:1 proportion into training and test tests. The training set was used for two rounds of cross-validation: first to optimise individually a number of possible machine learning algorithms, and second to select the best performing tool. (3) AdaBoost was the best performing algorithm, and forms the basis of CardioBoost. (4) CardioBoost was benchmarked against existing best-in-class tools using the hold-out test data, (5) a number of additional independent test sets, and (6) approaches based on association with clinical characteristics of variant carriers that do not rely on a gold-standard classification. (b) Illustrative distributions of predicted pathogenicity scores for a set of pathogenic and benign variants obtained by a hypothetical binary classifier. In a clinical context (based on ACMG/AMP guidelines), variants are classified into the following categories according to the probability of pathogenicity: disease-causing (Probability of pathogenicity (Pr) >=0.9), benign/likely benign (Pr <=0.1) and a clinically indeterminate group of Variants of Uncertain Significance with low interpretative confidence (0.1 < Pr < 0.9). (c)The corresponding confusion matrix with the defined double classification thresholds Pr >=0.9 and Pr <=0.1.

2.3.2 CardioBoost outperforms state-of-the-art genome-wide prediction tools based on overall classification performance measures

To estimate the classifiers' performance on variants of uncertain significance, I evaluated their classification performances on the hold-out test sets. CardioBoost was compared against state-of-the-art genome-wide variant pathogenicity predictors including M-CAP[39], REVEL[40], CADD[90], Eigen[42] and PrimateAI[15], reported to have leading performance in pathogenicity prediction of rare missense variants. Classification performance was first summarised using the area under the Precision-Recall Curve[111] (PR-AUC) and the area under the Receiver Operating Characteristic Curve (ROC-AUC), without relying on a single pre-defined classification threshold to discriminate disease-causing and benign variants.

In both inherited cardiac conditions, CardioBoost achieved the best values in both PR-AUC and ROC-AUC (**Figure 2.2**). The difference in performance was statistically significant for cardiomyopathies, with significantly increased PR-AUC (maximum *P*-value = 0.005 between the pairwise statistical comparisons using permutation test) and ROC-AUC (maximum *P*-value = 5×10^{-6} between the statistical comparisons using Delong test[112]). Among probabilistic predictors (CardioBoost, M-CAP, REVEL and PrimateAI), CardioBoost has significantly increased Brier score for both cardiomyopathies (maximum *P*-value = 0.005 between the pairwise comparisons via permutation test) and arrhythmia syndromes (maximum *P*-value = 0.02 between the pairwise comparisons via permutation test) (**Table 2.9**).

In the subsequent benchmarking studies, I specifically demonstrate CardioBoost performances compared with M-CAP and REVEL since they are explicitly trained to distinguish rare disease-causing variants from rare benign ones using ensemble learning approaches comparable to CardioBoost, and their overall classification performances are representative of these state-of-the-art tools shown in the above analysis. As the pathogenicity scores of M-CAP and REVEL were used as input features for CardioBoost, CardioBoost might indirectly expose to variants used in their previous training. This might worsen classification performance if the variants were erroneously classified during upstream training, or lead to inflated performance estimates through overfitting, so I also investigated the extent to which these potential limitations influenced CardioBoost performance. CardioBoost was shown to consistently improve on cardiomyopathy- and arrhythmia-specific prediction over existing genome-wide tools both on indirectly “seen” (variants used to train upstream genome-wide learners) and “unseen” (completely novel) data. The overall accuracy of CardioBoost between the unseen and seen datasets is not significantly different for either CM or IAS. (Results shown in **Table 2.10** and **Table 2.11;** Methods described in **2.2.11**).

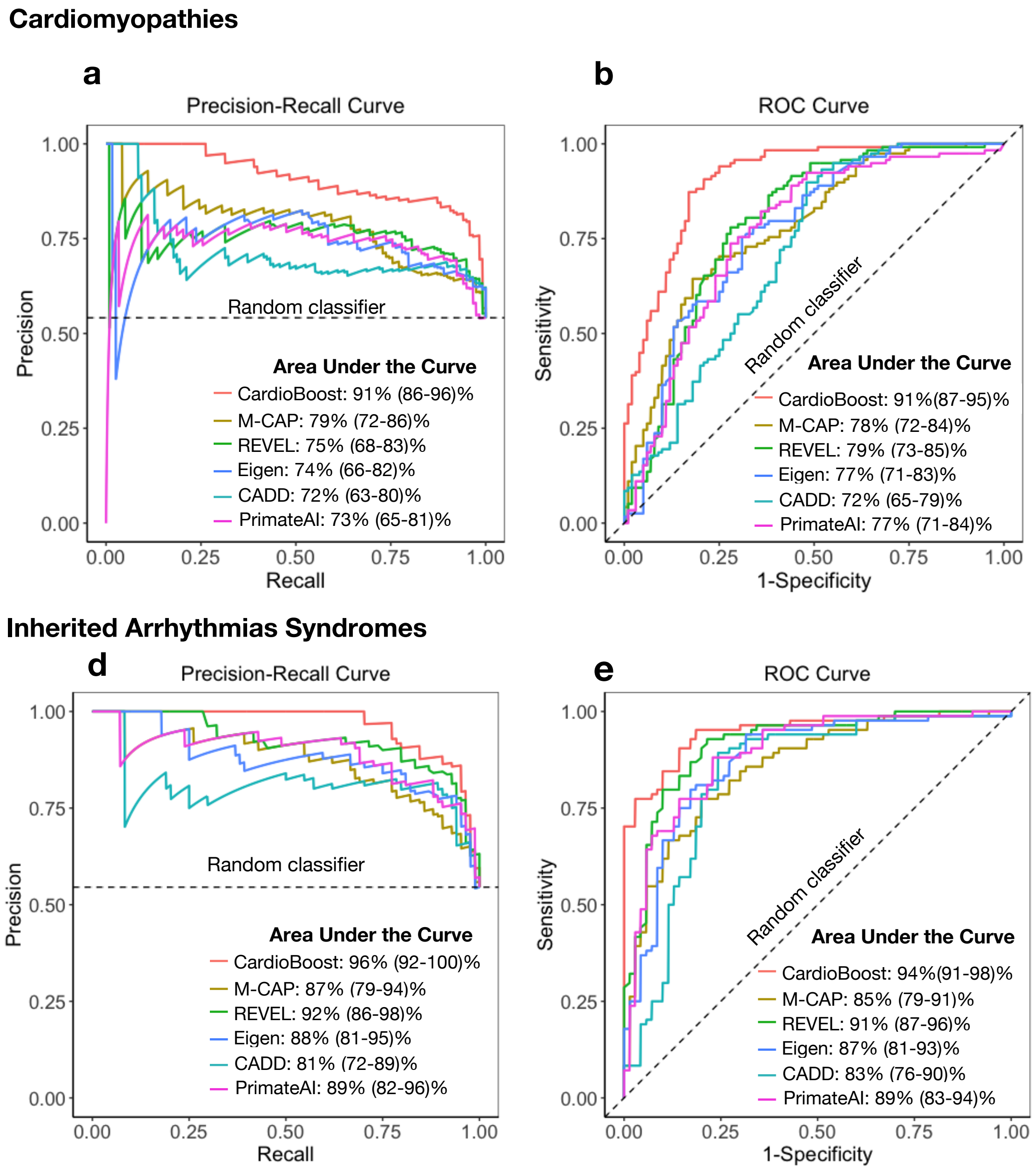


Figure 2.2 CardioBoost outperforms state-of-the art genome-wide prediction tools on hold-out test data. (**a-c**) Precision-Recall Curves, ROC Curves and Brier Scores for cardiomyopathy variant pathogenicity prediction. (**d-f**) Precision-Recall Curves, ROC Curves and Brier Score for inherited arrhythmia variant pathogenicity prediction. The dashed lines demonstrate the performance of a random classifier.

Table 2.9 Brier Scores to compare performances of probabilistic variant pathogenicity predictions in the hold-out test data set.

	Cardiomyopathies	Arrhythmias
CardioBoost	**0.12**	**0.09**
M-CAP	0.20	0.17
REVEL	0.19	0.17
PrimateAI	0.21	0.18

Table 2.10 Performance comparison on variants "unseen" and indirectly "seen" in the hold-out test data set for cardiomyopathy variant pathogenicity prediction. To assess whether bias is introduced in evaluating variants previously used in the training of M-CAP and REVEL, the performance of CardioBoost on wholly "unseen" data (not used in the training of M-CAP and REVEL), and indirectly "seen" data" (used in the training of M-CAP and REVEL) were compared with M-CAP and REVEL. For each predictive performance measure, the best algorithm is highlighted in bold.

	"Unseen" data $N_{pathogenic}$ = 41 N_{benign} = 24			"Seen" data $N_{pathogenic}$ = 77 N_{benign} = 76		
	CardioBoost (%)	M-CAP (%)	REVEL (%)	CardioBoost (%)	M-CAP (%)	REVEL (%)
PR-AUC	**90.2**	80.2	73.8	**91.8**	78.6	76.7
ROC-AUC	**86.3**	71.1	70.2	**92.1**	79.8	81.9
Brier Score	**13.4**	21.5	19.5	**11.8**	19.0	19.2
Overall Accuracy	**60.0**	30.8	12.3	**64.7**	27.5	19.6
Proportion of variants classified with high confidence	**69.2**	40.0	20.0	**70.6**	31.4	22.9
Accuracy of high-confidence classifications	**86.7**	76.9	61.5	**91.7**	87.5	85.7
Proportion of variants with indeterminate classifications	**30.8**	60.0	80.0	**29.4**	68.6	77.1
TPR	**70.7**	43.9	19.5	**68.8**	40.3	32.5
PPV	**82.9**	75.0	61.5	**88.3**	86.1	83.3
TNR	**41.7**	8.3	0.0	**60.5**	14.5	6.6
NPV	**100.0**	**100.0**	NA[1]	95.8	91.7	**100.0**

[1] No variants are classified as benign by REVEL.

Table 2.11 Performance comparison on variants "unseen" and indirectly "seen" in the hold-out test data set for arrhythmia variant pathogenicity prediction. To assess whether bias is introduced in evaluating variants previously used in the training of M-CAP and REVEL, the performance of CardioBoost on entirely "unseen" data (not used in the training of M-CAP and REVEL), and indirectly "seen" data" (used in the training of M-CAP and REVEL) were compared with M-CAP and REVEL. For each predictive performance measure, the best algorithm is highlighted in bold.

	"Unseen" data $N_{pathogenic}$ = 17 N_{benign} = 18			"Seen" data $N_{pathogenic}$ = 67 N_{benign} = 52		
	CardioBoost (%)	M-CAP (%)	REVEL (%)	CardioBoost (%)	M-CAP (%)	REVEL (%)
PR-AUC	**94.4**	82.2	87.1	96.8	88.6	93.1
ROC-AUC	**94.1**	85.6	86.3	95.0	84.6	92.6
Brier Score	**12.2**	15.9	20.6	**9.3**	17.4	16.2
Overall Accuracy	**80.0**	34.3	28.6	**81.5**	29.4	39.5
Proportion of variants classified with high confidence	**88.6**	40.0	34.3	**88.2**	31.9	42.0
Accuracy of high-confidence classifications	**90.3**	85.7	83.3	92.4	92.1	**94.0**
Proportion indeterminate classifications	**11.4**	60.0	65.7	**11.8**	68.1	58.0
TPR	**88.2**	70.6	58.8	**82.1**	43.3	67.2
PPV	**88.2**	85.7	83.3	91.7	93.5	**93.8**
TNR	72.2	0.0	0.0	**80.8**	11.5	3.8
NPV	**92.9**	NA	NA	93.3	85.7	**100.0**

2.3.3 CardioBoost outperforms existing genome-wide prediction tools on high-confidence classification measures

In addition to estimating conventional classification performance, I evaluated performance at thresholds corresponding to accepted levels of certainty required for clinical decision making[2] (90%; see definitions on **Figure 2.1**b). Using these thresholds (disease-causing: probability of pathogenicity (Pr) ≥ 0.9; benign/likely benign: Pr ≤ 0.1; indeterminate: 0.1 < Pr < 0.9), CardioBoost again outperforms existing genome-wide machine learning variant classification tools when assessed using hold-out test data (**Table 2.12**).

CardioBoost maximises the identification of both disease-causing and benign variants. In both conditions, CardioBoost had the highest true positive rate (TPR) (CM 69.5%; IAS 83.3%) and true negative rate (TNR) (CM 56%; IAS 78.6%) (**Table 2.12**, *P*-value < 0.001). In total, CardioBoost correctly classified 63.3% of cardiomyopathy test variants and 81.2% of arrhythmia test variants with 90% or greater confidence-level. The proportions of correctly classified variants are significantly higher (*P*-value < 0.001) than those obtained with M-CAP (CM 28.4%; IAS 30.5%) and REVEL (CM 17.4%; IAS 37%). In addition, CardioBoost minimises the number of indeterminate variants. Only 29.8% of cardiomyopathy test variants and 11.7% of arrhythmia test variants achieved indeterminate scores between 0.1 and 0.9, which were significantly fewer (*P*-value < 0.001) than those obtained with M-CAP (CM 66.1%; IAS 66.2%) or REVEL (CM 78%; IAS 59.7%) (**Table 2.12**).

Overall, using these thresholds CardioBoost assigned high-confidence classifications to 70.2% of cardiomyopathy test variants, among which 90.2% were correct. For arrhythmias, CardioBoost reported 88.3% of test variants with high confidence, with 91.9% prediction accuracy. The reported results are robust to the choice of classification thresholds. While guidelines propose 90% confidence as appropriate thresholds for likely pathogenic or likely benign classifications, some may advocate a higher confidence threshold. When assessed at a 95%-certainty classification threshold, CardioBoost continues to consistently outperform genome-wide tools with significantly (*P*-value < 0.001) higher accuracies (**Table 2.13**).

Table 2.12 CardioBoost outperforms existing genome-wide tools for the classification of hold-out test variants. CardioBoost outperforms existing genome-wide machine learning tools for the classification of hold-out test variants. The performance of each tool is reported using the high-confidence variant classification thresholds: high-confidence disease-causing (Pr ≥ 0.9), high-confidence benign (Pr ≤ 0.1), and indeterminate. For each predictive performance measure, the best algorithm is highlighted in bold. Permutation tests were performed to evaluate whether the performance of CardioBoost was significantly different from the best value obtained by M-CAP or REVEL (significance levels: ***P-value ≤ 0.001, **P-value ≤ 0.01, *P-value ≤ 0.05).

(%)	Cardiomyopathies			Arrhythmias		
	CardioBoost	M-CAP	REVEL	CardioBoost	M-CAP	REVEL
Overall accuracy	**63.3*** **	28.4	17.4	**81.2*** **	30.5	37
Proportion of variants classified with high confidence	**70.2*** **	33.9	22	**88.3*** **	33.8	40.3
Accuracy of high-confidence classifications	**90.2**	83.8	79.2	**91.9**	90.4	**91.9**
Proportion of variants with indeterminate classification	**29.8*** **	66.1	78	**11.7*** **	66.2	59.7
TPR	**69.5*** **	41.5	28	**83.3*** **	48.8	65.5
PPV	**86.3**	81.7	76.7	90.9	91.1	**91.7**
TNR	**56*** **	13	5	**78.6*** **	8.6	2.9
NPV	96.6	92.9	**100**	93.2	85.7	**100**

Table 2.13 CardioBoost outperforms existing genome-wide classification tools for the classification of hold-out test variants using 95%-certainty thresholds. While 90% is defined as a high-confidence threshold for clinical action in the ACMG/AMP guidelines, some may advocate a more stringent approach. I therefore assessed the performance of each tool using more stringent values for clinically relevant variant classification thresholds: high-confidence disease-causing (Pr ≥ 0.95), high-confidence benign (Pr ≤ 0.05), and indeterminate. For each predictive performance measure, the best algorithm is highlighted in bold. Permutation tests were performed to evaluate whether the performance of CardioBoost was significantly different from the best value obtained by M-CAP or REVEL (significance levels: ****P*-value ≤ 0.001, ***P*-value ≤ 0.01, **P*-value ≤ 0.05).

(%)	Cardiomyopathies			Arrhythmias		
	CardioBoost	M-CAP	REVEL	CardioBoost	M-CAP	REVEL
Overall accuracy	**54.6*****	16.5	7.3	**78.6*****	7.8	22.1
Proportion of variants classified with high confidence	**60.1*****	18.8	10.1	**85.1*****	8.4	23.4
Accuracy of high confidence classifications	**90.8**	87.8	72.7	92.4	92.3	**94.4**
Proportion of variants with indeterminate classification	**39.9*****	81.2	89.9	**14.9*****	91.6	76.6
TPR	**62.7*****	24.6	11.9	**79.8*****	11.9	39.3
PPV	**87.1**	85.3	70.0	91.8	90.9	**93.9**
TNR	**45.0*****	7.0	2.0	**77.1*****	2.9	1.4
NPV	97.8	**100.0**	**100.0**	93.1	**100.0**	**100.0**

CardioBoost is not intended to replace a full expert variant assessment in clinical practice, but for comparative purposes it is informative to consider how classification performance changes under different application contexts. PPV and NPV are both dependent on the proportion of pathogenic variants in the variant set being tested, and so it is important to consider how our benchmarking translates to real-world application. Here I used the TPR, and TNR calculated on the hold-out test set to derive estimates of PPV and NPV for CardioBoost applied in different contexts where the true proportion of pathogenic variants might differ. My estimation provides a lower bound of PPV and NPV under the assumption that pathogenic variants are fully penetrant. In predictive genetic testing, the limitation of false positive prediction is prioritised, necessitating conservative estimates of PPV. Here I estimate reasonably conservative PPVs and corresponding NPVs of CardioBoost applied in two scenarios: in a diagnostic referral series and in samples from a general population. In a diagnostic laboratory cardiomyopathy referral series, where I estimate approximately 60% rare missense variants found in cardiomyopathy-associated genes to be pathogenic, the PPV and NPV of CardioBoost were estimated at 89% and 96% respectively. By contrast, in a general population, where I estimate the proportion of rare pathogenic variants of these ICC genes are ~ 1%, the PPV and NPV reach 5% and 99.9%. Similarly, I estimated the performance of CardioBoost in an arrhythmia cohort (PPV: 95%; NPV: 87%) and a general population (PPV:3%; NPV: 99.9%). This suggests that the predictions of disease-causing variants by CardioBoost are calibrated for high confidence only when applied in a diagnostic context, as would be expected. Classifications are appropriate for variants found in patients, with a reasonable prior probability of pathogenicity (details are described in Estimating the proportion of pathogenic missense variants in a diagnostic series and a general population).

Finally, as novel pathogenic variants are more likely to be ultra-rare (Minor allele frequency < 0.01%), I also tested CardioBoost performance on a hold-out set of only ultra-rare variants and confirmed that it consistently outperforms existing genome-wide tools (**Table 2.14**). Its performance on ultra-rare variants is comparable with that on rare variants.

Table 2.14 Comparison of classification performance on the hold-out test data set with minor allele frequency < 0.01%. As novel pathogenic variants are more likely to be ultra-rare, CardioBoost was tested on the hold-out set of only ultra-rare variants and was confirmed to have comparable performance with that on rare variants. The performance of each tool is reported using the 90% high-confidence variant classification thresholds: high confidence disease-causing (Pr ≥ 0.90), high confidence benign (Pr ≤ 0.10), and indeterminate. For each predictive performance measure, the best algorithm is highlighted in bold. Permutation tests were performed to evaluate whether the performance of CardioBoost was significantly different from the best value obtained by M-CAP or REVEL (significance levels: ***P-value ≤ 0.001, **P-value ≤ 0.01, *P-value ≤ 0.05).

	Cardiomyopathies			Arrhythmias		
	CardioBoost (%)	M-CAP (%)	REVEL (%)	CardioBoost (%)	M-CAP (%)	REVEL (%)
Classification performance measures						
PR-AUC	**93***	85	81	**97**	90	95
ROC-AUC	**91*****	79	79	**95**	86	93
Brier Score	**0.11***	0.18	0.17	**0.09**	0.15	0.14
90% high-confidence classification performance measures						
Overall accuracy	**64.9*****	30.9	19.7	**83.6*****	33.6	42.5
Proportion of variants classified with high confidence	**71.3*****	35.6	22.9	**93.3*****	93.8	95
Accuracy of high confidence classifications	**91.0**	86.6	86	93.3	93.8	**95**
Proportion of variants with indeterminate classification	**28.7*****	64.4	77.1	**6.7*****	65.2	53.3
TPR	**70.1*****	41.9	28.2	**85.4*****	50	67.1
PPV	**89.1**	86	84.6	94.6	**95.3**	94.8
TNR	**56.3*****	12.7	5.6	**80.8*****	7.7	3.8
NPV	95.2	90	**100**	91.3	80	**100**

2.3.4 Replication on additional independent test data confirms that CardioBoost improves prediction of disease-causing and benign variants

I collected four additional sets of independent test data to further assess the CardioBoost performance, using variants reported as pathogenic in ClinVar and HGMD[113] (both databases of aggregated classified variants), a diagnostic laboratory referral series from the Oxford Molecular Genetics Laboratory (OMGL), and a large registry of HCM patients, SHaRe[57]. When using ClinVar variants to test CM, only variants with two-star review status (i.e. criteria provided, multiple submitters, no conflicts) are included. CardioBoost consistently achieved the highest TPRs: predicting the most disease-causing variants with over 90% certainty (**Table 2.15**). On a set of rare variants found in the gnomAD reference dataset, which is not enriched for inherited cardiac conditions and hence where the prevalence of disease should be equivalent to the general population, CardioBoost consistently predicts the most variants as benign (**Table 2.15**). I also assessed the accuracy of CardioBoost using cell-based functional mapping of amino acid substitutions in calmodulin genes (*CALM1*, *CALM2* and *CALM3*) from a previous deep mutational scanning (DMS) study[75]. Averaged over three calmodulin genes, CardioBoost has the significantly highest accuracy to predict the DMS classification (**Table 2.15**). CardioBoost also performed the best when assessed at a higher 95%-certainty classification threshold (**Table 2.16**) and on sets of ultra-rare variants (**Table 2.17**).

Table 2.15 Evaluation of performance on additional test sets. CardioBoost performance was evaluated against additional variant sets. Four resources provided known pathogenic variants (SHaRe cardiomyopathy registry, ClinVar (two-star submissions), a UK regional genetic laboratory (Oxford Medical Genetics Laboratory – OMGL) and the Human Gene Mutation Database – HGMD). Variants found in gnomAD population controls were expected to be predominantly benign. Since gnomAD includes variants seen in the previous ExAC dataset that was partially used to train M-CAP and REVEL, I tested against the subset of variants in gnomAD that were not in ExAC. The number of single nucleotide variants in each set is shown in brackets. I also evaluated the classification accuracies on functional mapping of amino acid substitutions in calmodulin genes obtained through a previous deep functional scanning study.

	Cardiomyopathies			
	Pathogenic test variants (TPR)			Benign/population test variants (TNR)
	SHaRe (N = 129)	ClinVar (N = 15)	HGMD (N = 145)	gnomAD (N = 2,003)
CardioBoost	**62.0*****	**66.7**	**41.4*****	**51.5*****
M-CAP	37.2	40.0	22.1	20.3
REVEL	24.0	53.3	22.8	5.6

	Arrhythmias			
	Pathogenic test variants (TPR)		Benign/Population test variants (TNR)	Deep Mutational Scanning (Accuracy)
	OMGL (N = 77)	HGMD (N = 138)	gnomAD (N = 1,237)	Calmodulin (N = 576)
CardioBoost	**88.3*****	**72.5*****	**64.3*****	**29.0*****
M-CAP	59.7	39.9	9.8	0.3
REVEL	68.8	52.9	2.8	4.2

Table 2.16 Evaluation of performances on additional test sets using 95%-certainty threshold.

(%)	Cardiomyopathies			
	Pathogenic test variants (TPR)			Benign test variants (TNR)
	SHaRe (N = 129)	ClinVar (N = 15)	HGMD (N = 145)	gnomAD (N = 2,003)
CardioBoost	**51.2*****	**60.0***	**33.8*****	**44.2*****
M-CAP	19.4	13.3	9.0	9.9
REVEL	6.2	6.7	6.9	2.6
	Arrhythmias			
	Pathogenic test variants (TPR)		Benign test variants (TNR)	Deep Mutational Scanning (Accuracy)
	OMGL (N = 77)	HGMD (N = 138)	gnomAD (N = 1,237)	Calmodulin (N = 576)
CardioBoost	**87.0*****	**71.0*****	**61.3*****	**25.7*****
M-CAP	23.4	18.8	4.3	0
REVEL	28.6	23.9	1.2	0.3

Table 2.17 Evaluation of performances on additional test sets with minor allele frequency < 0.01%.

(significance levels: ****P*-value ≤ 0.001, ***P*-value ≤ 0.01, **P*-value ≤ 0.05).

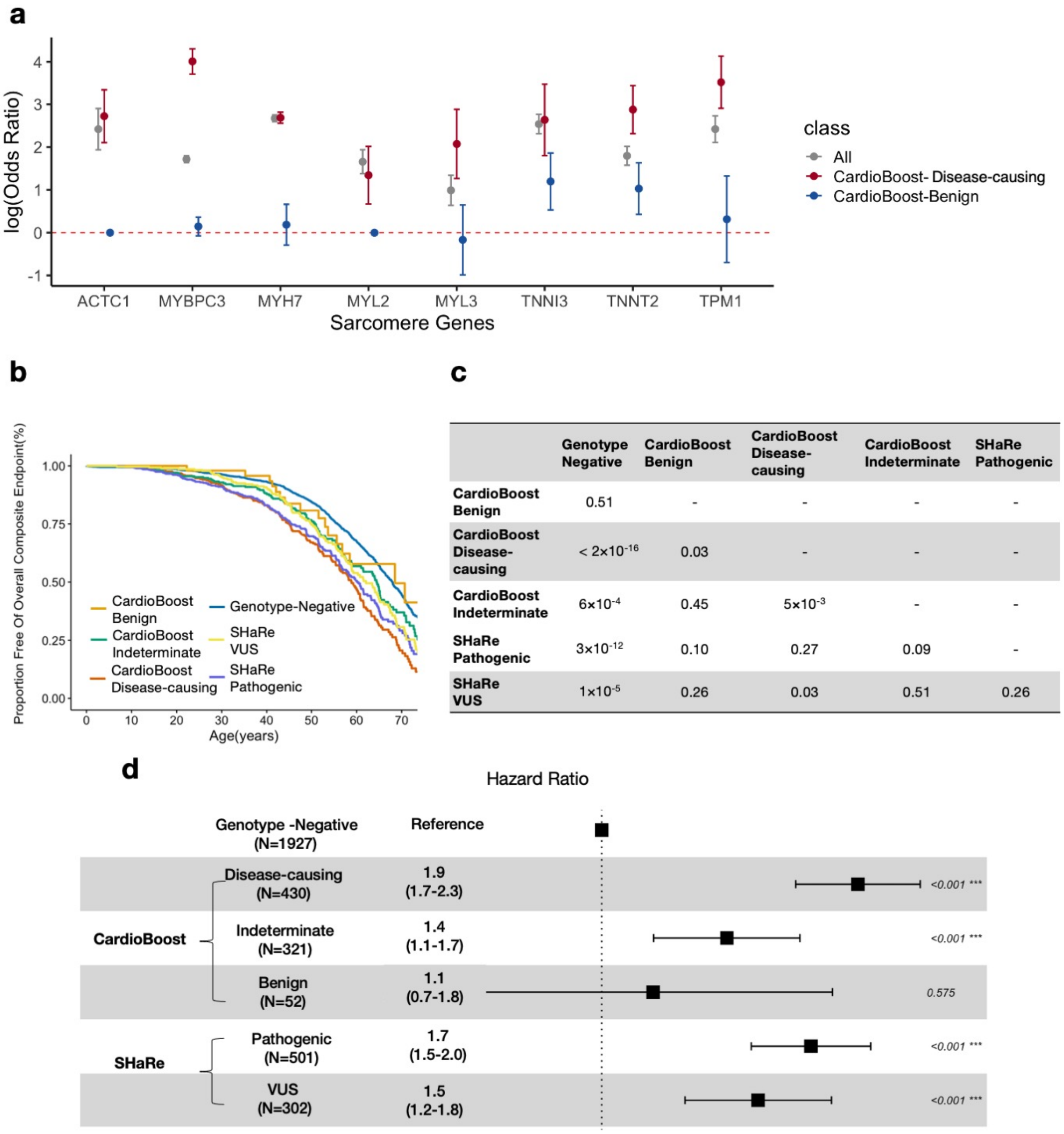
(%)	Cardiomyopathies			
	Pathogenic test variants (TPR)			Benign test variants (TNR)
	SHaRe (N = 129)	ClinVar (N = 14)	HGMD (N = 143)	gnomAD (N = 1,999)
CardioBoost	**62.0*****	**71.4***	**42.0*****	**51.5*****
M-CAP	37.2	42.9	22.4	20.3
REVEL	24.0	57.1	23.1	5.7
	Arrhythmias			
	Pathogenic test variants (TPR)		Benign test variants (TNR)	Deep Mutational Scanning (Accuracy)
	OMGL (N = 77)	HGMD (N = 138)	gnomAD (N = 1,232)	Calmodulin (N = 576)
CardioBoost	**88.3*****	**72.5*****	**64.4*****	**29.0*****
M-CAP	59.7	39.9	9.8	0.3
REVEL	68.8	52.9	2.8	4.2

2.3.5 CardioBoost discriminates variants that are highly disease associated

Since benchmarking against a gold-standard variant set may be susceptible to classification errors in the data, I employed two additional approaches to evaluate CardioBoost predictions directly against patient characteristics, to confirm biological and clinical relevance.

First, I directly assessed the strength of the association between the specified disease and rare variants stratified by the different tools. I compared the proportions of rare missense variants in a cohort of 6,327 genetically-characterised patients with HCM, from the SHaRe registry[57], with 138,632 reference samples from gnomAD v2.0 (**Figure 2.3a**). I calculated the Odds Ratio (OR) for all rare variants observed in each sarcomere gene, and for variants stratified by CardioBoost, M-CAP, and REVEL after excluding variants seen in our training data.

For seven out of the eight CM-associated genes (*MYH7*,*TNNI3*,*TPM1*,*ACTC1*,*TNNT2*, *MYBPC3* and *MYL3*), the OR for variants prioritised by CardioBoost (i.e. predicted disease-causing with Pr ≥ 0.9) was greater than the baseline OR (including all observed variants without discriminating disease-causing and benign variants), indicating that the tool is discriminating a set of variants more strongly associated with the disease. For three genes (*TPM1*, *TNNT2*, *MYBPC3*), the difference was statistically significant (*P*-value < 0.05). Concordantly, variants in seven out of the eight sarcomere genes predicted as benign have significantly decreased association with the disease compared with the baseline OR (*P*-value < 0.05). By contrast, M-CAP or REVEL did not show any demonstrable difference in disease ORs between predicted disease-causing and predicted benign variants (**Table 2.18**).



	Genotype Negative	CardioBoost Benign	CardioBoost Disease-causing	CardioBoost Indeterminate	SHaRe Pathogenic
CardioBoost Benign	0.51	-	-	-	-
CardioBoost Disease-causing	$< 2\times10^{-16}$	0.03	-	-	-
CardioBoost Indeterminate	6×10^{-4}	0.45	5×10^{-3}	-	-
SHaRe Pathogenic	3×10^{-12}	0.10	0.27	0.09	-
SHaRe VUS	1×10^{-5}	0.26	0.03	0.51	0.26

Figure 2.3 CardioBoost improves prioritisation of variants associated with disease and clinical outcomes in patients with HCM. (**a**) The ORs (on log scale) for three groups of variants were compared: (i) all rare variants, (ii) rare variants predicted disease-causing by CardioBoost (Pr ≥ 0.9, and excluding those seen in our training data), and (iii) rare variants predicted as benign by CardioBoost (Pr ≤ 0.1 and excluding those seen in our training data). For most of the sarcomere-encoding genes, variants classified as disease-causing by CardioBoost are enriched for disease-association, and those classified as benign are depleted, compared with unstratified rare missense variants. (**b-d**) CardioBoost variant

classification stratifies key clinical outcomes in patients with HCM. Clinical outcomes provide an opportunity to assess classifier performance independent of the labels used in the gold-standard training data. (**b**) Kaplan-Meier event-free survival curves are shown for patients in the SHaRe cardiomyopathy registry, stratified by genotype as interpreted by CardioBoost. The patients carrying variants seen in the CardioBoost training set were excluded from this analysis. Patients with predicted disease-causing variants in sarcomere-encoding genes have more adverse clinical events compared with patients without sarcomere-encoding variants ("genotype-negative"), and compared with patients with sarcomere-encoding variants classified as benign. Survival curves stratified by variants as adjudicated by experts (marked in figure with prefix "SHaRe") are shown for comparison. The composite endpoint comprised the first incidence of any component of the ventricular arrhythmic or heart failure composite endpoint, atrial fibrillation, stroke, or death. (**c**) *P*-values of the log-rank test in the pairwise comparisons of Kaplan-Meier survival curves. (**d**) Forest plot displays the hazard ratio (with confidence interval) and *P*-value of tests comparing patients' survival stratified by CardioBoost classification and SHaRe experts' classification based on Cox proportional hazards models.

Table 2.18 CardioBoost variant classification stratifies variants with increased disease Odds Ratio for sarcomere-encoding genes. Odd Ratios (ORs) and their confidence intervals were calculated for rare variants observed in sarcomere-encoding genes using SHaRe HCM cohorts and gnomAD. The ORs for three groups of variants were compared: (i) all rare variants, (ii) rare variants predicted disease-causing by CardioBoost (Pr ≥ 0.9, and excluding those seen in our training data), and (iii) rare variants predicted as benign by CardioBoost (Pr ≤ 0.1 and excluding those seen in our training data). The ORs of variants classified by M-CAP and REVEL were also calculated.

Gene symbol	all observed rare variants (95% CI)	CardioBoost disease-causing variants (95% CI)	CardioBoost benign variants (95% CI)	M-CAP disease-causing variants (95% CI)	M-CAP benign variants (95% CI)	REVEL disease-causing variants (95% CI)	REVEL benign variants (95% CI)
MYH7	14.5 (13.4-15.7)	14.7 (12.9-16.7)	1.2 (0.7-1.9)	14.8 (12.9-16.9)	-[1]	15.9 (13.1-19.2)	-[1]
TNNI3	12.6 (10.1-15.9)	14.0 (6.1-32.3)	3.3 (1.7-6.4)	1.0 (1 -1.1)	4.7 (1.6 – 14)	12.1 (4-35.9)	1.0 (1-1.1)
TPM1	11.2 (8.2-15.3)	33.7 (18.3 – 62.2)	1.4 (0.5-3.8)	1.0 (1 -1.1)	0.5 (0.1 – 3.6)	38.9 (5.9-256.6)	-[1]
ACTC1	11.2 (6.9-18.2)	15.2 (8.2-28.3)	1.0 (1-1.1)	1.0 (1 -1.1)	1.0 (1 - 1.1)	19.8 (9.4-42)	-[1]
TNNT2	6.0 (4.8-7.5)	17.7 (10.1-31.1)	2.8 (1.5-5.1)	1.0 (1 -1.1)	1.0 (0.1 – 7.1)	25.8 (3.3-199.1)	28.9 (5.2-161.6)
MYBPC3	5.6 (5.1-6.0)	55.1 (41-74.1)	1.2 (0.9-1.4)	1.0 (1 -1.1)	0.7 (0.4-1.1)	12.8 (7.6-21.8)	1.2 (0.8-1.8)
MYL2	5.2 (4.0-6.9)	3.8 (2.0-7.5)	1.0 (0.9-1.1)	1.0 (1 -1.1)	0.2 (0-1.6)	1.7 (0.4-7)	1.0 (1-1.1)
MYL3	2.7 (1.9-3.8)	7.9 (3.5-17.8)	0.8 (0.4-1.9)	1.0 (1 -1.1)	0.3 (0-2.2)	19.4 (8.3-45.4)	-[1]

[1] OR not calculated since the number of missense variants predicted as benign is zero in the gnomAD population.

2.3.6 CardioBoost variant classification is associated with adverse clinical outcome

As a further assessment independent of gold-standard classification, I tested the association of variants stratified by CardioBoost with clinical outcomes in the same cohort of patients. Patients with HCM who carry known pathogenic variants in genes encoding sarcomeric proteins have been shown to follow an adverse clinical course compared with "genotype-negative" individuals (no rare pathogenic variant or VUS in a sarcomere-encoding gene, and no other pathogenic variant identified)[57,114,115], with a higher burden of adverse events. Patients carrying benign variants in HCM-associated genes would be expected to follow a similar trajectory to those genotype-negative patients.

I evaluated clinical outcomes in a subset of the SHaRe cohort comprising of 803 HCM patients each with a rare missense variant in a sarcomere-encoding gene, and 1,927 genotype-negative HCM patients, after excluding all patients carrying variants that were seen in the CardioBoost training set. I compared event-free survival (i.e. age until the first occurrence of a composite adverse clinical outcome including heart failure events, arrhythmic events, stroke, and death) of these patients, stratified by CardioBoost-predicted pathogenicity (the full definition of a composite adverse clinical outcome is described in Survival analysis).

CardioBoost classification stratifies novel variants with significantly different patient-survival curves (**Figure 2.3b-Figure 2.3d**). Patients carrying variants predicted as disease-causing (CardioBoost disease-causing) were likely to have earlier onset and a higher adverse event rate than those without identified rare variants (CardioBoost disease-causing vs genotype negative: *P*-value < 2×10^{-16}; Hazard Ratio (HR) = 1.9), or those with variants predicted to be benign (CardioBoost disease-causing vs CardioBoost benign: *P*-value = 0.03; HR = 1.7). The probability of developing the overall composite outcome by age 60 is 54% (95% CI: 46%-59%) for CardioBoost disease-causing patients, versus 33% (95% CI: 30%-35%) for genotype-negative patients. By contrast, groups stratified by M-CAP or REVEL variant classification did not show significantly different event-free survival time (M-CAP disease-causing vs M-CAP benign: *P*-value = 0.31; REVEL disease-causing vs REVEL benign: *P*-value = 0.30) (**Figure 2.4**).

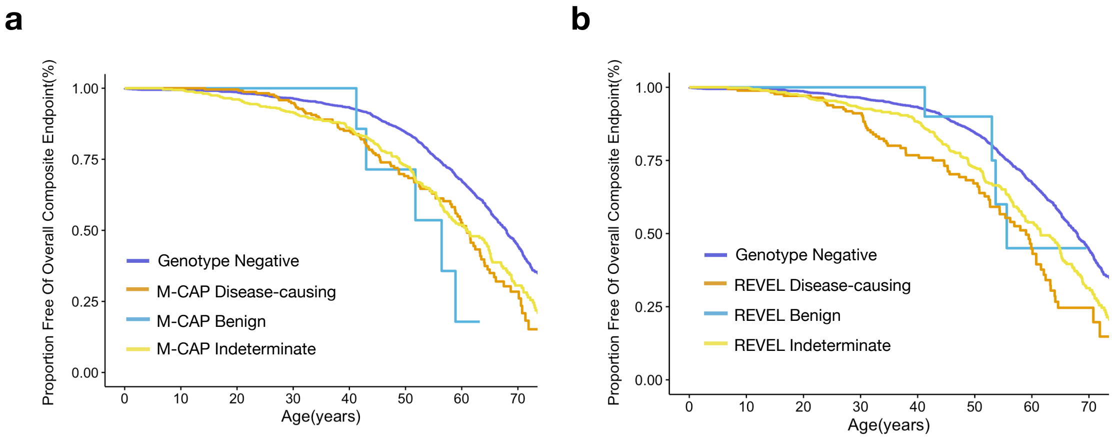


Figure 2.4 Variants classification by state-of-the-art genome-wide tools M-CAP and REVEL did not show to stratify the survival outcomes of patients. (**a**) Kaplan-Meier event-free survival curves for patients in the SHaRe cardiomyopathy registry, stratified by genotype as interpreted by M-CAP. The patients with variants predicted disease-causing by M-CAP did not have significantly different survival time compared to those with predicted benign variants (log-rank test *P*-value = 0.31). (**b**) Kaplan-Meier event-free survival curves for patients in the SHaRe cardiomyopathy registry, stratified by genotype as interpreted by REVEL. Patients with predicted disease-causing variants by REVEL did not have significantly different survival time compared to those with predicted benign variants (log-rank test *P*-value = 0.30).

2.4 Discussion

The above results show that *in silico* prediction of variant pathogenicity for inherited cardiac conditions is improved within a disease-specific framework trained using expert-curated interpreted variants. This is demonstrated through improved classification performance, stronger disease association, and significantly improved stratification of patient outcomes over published genome-wide tools.

2.4.1 Strengths of the study

There are several factors that may contribute to improved performance for a gene- and disease-specific classifier like CardioBoost over genome-wide tools. First, the use of disease-specific labels could decrease the false prediction of benign variants as disease-causing. A variant causative of one Mendelian dominant disorder may be benign with respect to a different disorder (associated with the same gene), if the conditions result from distinct molecular pathways. Since genome-wide tools are trained on universal labels (i.e. whether a variant ever causes any diseases), they would be expected to yield false positive predictions in the context of specific diseases. Second, while the representative genome-wide tools M-CAP and REVEL are trained on variants from HGMD curated from literature, CardioBoost is trained on high-quality expert-curated variants, thus reducing label bias and increasing the prediction performances. Thirdly, as the genome-wide tools are trained across the genome, the learning function that maps the input features into the pathogenicity score is fitted using the training samples from all genes in the genome. However, different genes may have different mapping functions, for example related to different molecular mechanisms. Restricting to a set of well-defined disease-related genes may exclude influences from other unrelated genes.

Taking the hypothesis further, one might expect a gene-disease-specific model would be the most accurate since it represents the exact genotype-phenotype relationship. However, there is a trade-off between the size of available training data and the specialization of classification tasks. Here, CardioBoost groups together genes for two sets of closely related disorders, including three genes in which variants with different functional consequences lead to distinct phenotypes in our training set (i.e. *SCN5A*, *TNNI3, MYH7*). This is a potential limitation since distinct functional consequences might optimally be modelled separately or distinctly. I explored alternative models for cardiomyopathy classifiers, for which our training data set is larger than that for arrhythmias. Two disease-specific models (HCM-specific and DCM-specific) and three gene-syndrome-specific models (*MYH7*-HCM-specific, *MYH7*-DCM-specific, and *MYBPC3*-HCM-specific) with the largest training data size were built and compared (**Table 2.19**). None of the alternative models had comparable performance to the combined-cardiomyopathy model. Therefore, I conclude that given the current availability of training data, a cardiomyopathy-specific classifier provides the best empirical balance between grouping variants with similar phenotypic effects and making use of a relatively large training data set. It improves prediction both over genome-wide models that entirely ignore variants' phenotypic effects, and over gene-disease-specific models for which there is insufficient training data. Therefore, I adopted the broadly disease-specific models as our final classifier, but anticipate that complete separation of distinct phenotypes may be advantageous when more training data becomes available in the future.

Table 2.19 Comparison of out-of-sample classification performances for alternative disease-specific classification tasks. I explored alternative variant classification models as exemplified for cardiomyopathies with relatively larger size of training data: two syndrome-specific models (HCM-specific and DCM-specific) and three gene-syndrome-specific models (*MYH7*-HCM-specific, *MYH7*-DCM-specific, and *MYBPC3*-HCM-specific). Here the broadly cardiomyopathies-specific model was chosen since none of the alternative models had comparable performances.

Predictive task	Number of training variants	Precision-Recall AUC (%)
CM-specific	440	91
HCM-specific	348	79
DCM-specific	309	48
MYH7-HCM-specific	152	87
MYH7-DCM-specific	152	35
MYBPC3-HCM-specific	106	76

As another advantage of CardioBoost, it natively outputs a continuous probability of pathogenicity that is directly interpretable. Users may therefore define their own confidence thresholds according to the intended application. For example, users might want to use a lower probability threshold if they would like to prioritise sensitivity when the cost of false positive is neglectable. A posterior probability of variant pathogenicity could also be derived by incorporating our prediction score with further evidence, such as linkage scores calculated from the evaluation of segregation in a family.

While I have extensively benchmarked CardioBoost with genome-wide tools, the idea of gene-specific or syndrome-specific models for inherited cardiac conditions has been developed previously including a MYH7-specific predictor[116], a Bayesian syndrome-specific classification predictor APPRAISE[69], a HCM-specific classification model PolyPhen-HCM[117] and a cardiomyopathy-specific model PathoPredictor[118]. Compared to these existing important works, CardioBoost has improved the disease-specific classifiers in terms of the size and diversity of the predictive features and training datasets. I collected substantially relevant features (n=76) for variant classifications including conservation, existing pathogenicity scores and genetic constraint scores. It was also trained with larger size of high-quality expert-curated variants including as many disease genes as possible (CM: genes = 16, variants = 440; IAS: genes = 7, variants = 326).

2.4.2 Limitations of the study

There are several potential limitations and avenues for future refinement. First, I have only considered the prediction of pathogenicity for missense variants thus far. The inclusion of different classes of variants in disease-specific models is challenging since there is limited high-confidence training data for non-missense variants.

A second key limitation of CardioBoost is that it does not consider all relevant lines of evidence, and therefore it is not intended to serve as a tool for a comprehensive assessment of variant pathogenicity comparable to clinicians' interpretation based on ACMG guidelines. Some evidence types are limited by availability such as population allele frequency data and segregation data. Others could not be systematically included in a machine learning framework either because they are not well structured as in the case of functional data, *de novo* data, and allelic data, or they are too sparse. For example, many variants lack experimental data, and the precise population allele frequency of many variants is unknown, though this implies significant rarity. In our training data, 45% of variants in cardiomyopathies and 44% of variants in arrhythmias were not seen in the gnomAD control population. Here, I do not include allele frequencies in gnomAD as a predictive feature since the relation between variant pathogenicity and allele frequency scale beyond current observation is clearly unknown.

For these reasons, while the advantages of the proposed model are shown for variant classification in known disease genes over existing genome-wide tools, it’s necessary to emphasize to users that CardioBoost is not intended for use as a standalone clinical decision tool, or as a replacement for the existing ACMG/AMP guidelines for variant interpretation. Rather, in its current form it could provide a numerical value for evidence PP3 (“Multiple lines of computational evidence support a deleterious effect on the gene/gene product”) and BP4 (“Multiple lines of computational evidence suggest no impact on gene /gene product”) that is more reliable and accurate than existing genome-wide variant classifiers in the context of inherited cardiac conditions. High-confidence classifications by CardioBoost might appropriately activate PP3 (Pr>0.9) and BP4 (Pr<0.1). It is interpreted as the supporting evidence being activated with at least 90% confidence.

The widely-adopted ACMG/AMP framework is semi-quantitative, but one limitation is that the weightings applied to different rules are not all evidence-based or proven to be mathematically well-calibrated. It is anticipated that, with more training data and robust validation, quantitative tools like CardioBoost could be updated with more relevant lines of evidence adopting the principle of ACMG guidelines and will carry more weight in a quantitative decision framework than the current ACMG/AMP PP3 and BP4 rule affords.

While CardioBoost improves on existing tools, there remain a substantial number of variants receiving indeterminate classification by CardioBoost at high-confidence classification thresholds (**Table 2.12**: CM 29.8% IAS 11.7%). I anticipate that additional relevant functional annotations, accumulation of further gold-standard interpreted data and development of novel task-specific prediction approaches will continue to improve *in silico* prediction over time.

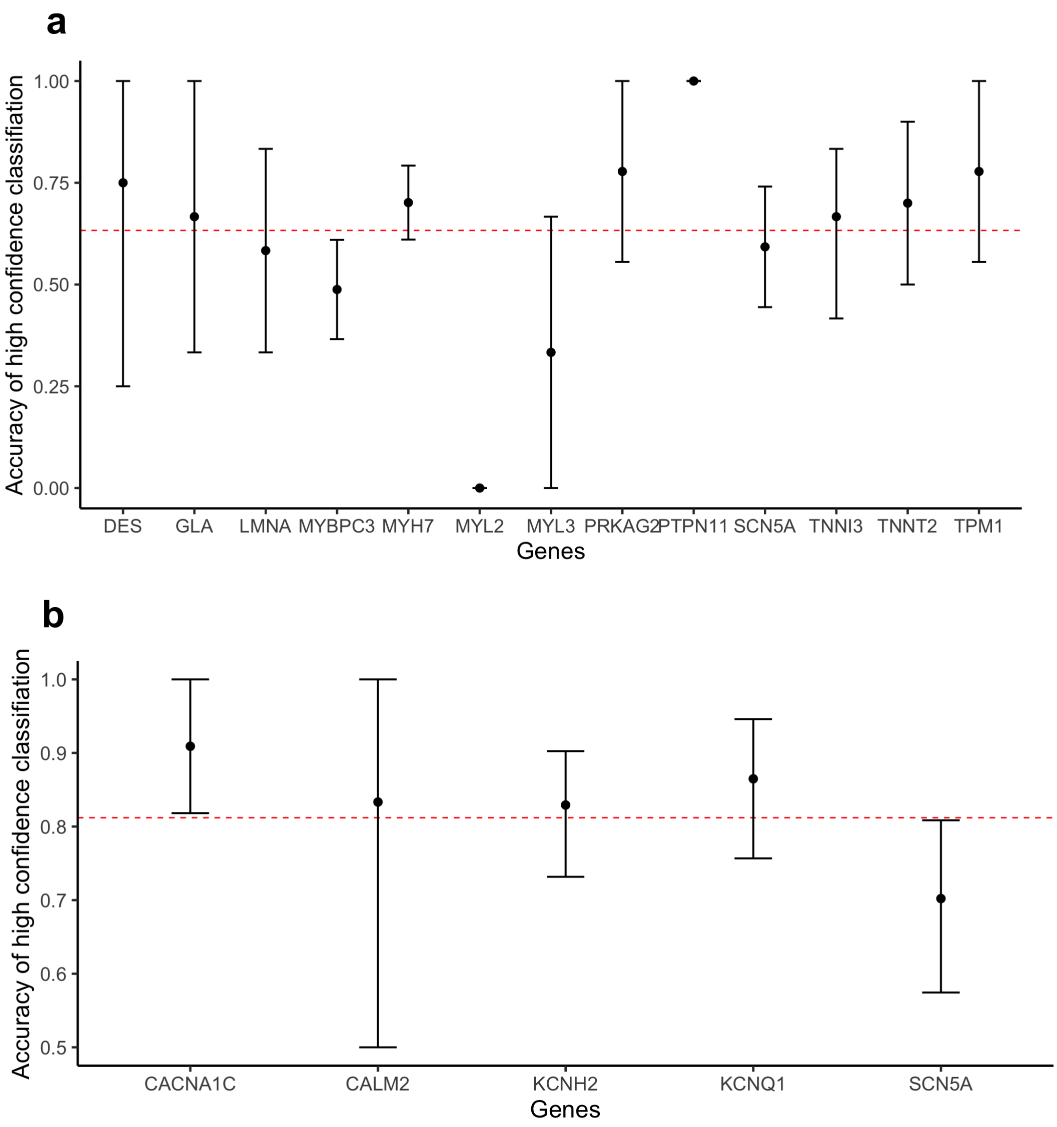


Figure 2.5 Variant classification performance per gene. The accuracy of high-confidence classification and its 90% bootstrap CI (n=1,000 times) are calculated per gene for (a) cardiomyopathies and (b) arrhythmias. The red dashed lines indicate the overall accuracies of variant classification at disease-level (extracted from Table 1). To be noticed, here the bootstrap CI is subjected to the size of test variants for each gene. Only genes with more than one test variants are considered in the analysis. Particular care should be taken for genes with wider confidence intervals in using CardioBoost for variant classification.

While CardioBoost performs well overall, the prediction performance and confidence differ by different genes according to the size of the training/test set for that gene. Five genes account for the majority of genetically explained cardiomyopathy and long QT (MYH7, MYBPC3, KCNQ1, KCNH2, SCN5A), resulting in narrower prediction confidence intervals. For other genes, the gold-standard data remain relatively sparse (**Figure 2.5**), resulting in wider prediction confidence intervals. Classifications of variants in these genes should be considered with appropriate care.

2.4.3 Conclusion

In conclusion, as exemplified in inherited cardiac conditions, I have substantiated that a disease-specific variant classifier improves the *in silico* prediction of variant pathogenicity over the best-performing genome-wide tools. This study also emphasizes the pitfalls of relying on genome-wide variant classifiers and the necessity to develop disease-specific variant classifiers to accurately interpret variant pathogenicity on specific phenotypes and diseases. I also highlight the need to evaluate variant pathogenicity prediction in clinical settings including accuracies on high confidence classification thresholds equivalent to accepted certainty required for clinical decision making, variants' association with disease and patients' clinical outcomes. To support accurate variant interpretation in inherited cardiac conditions, I provide pre-computed pathogenicity scores for all possible rare missense variants in genes associated with inherited cardiomyopathies and arrhythmias (https://www.cardiodb.org/cardioboost/). The demonstrated development and evaluation framework could be applicable to develop accurate disease-specific variant classifiers and improve variant interpretation in a wide range of Mendelian disorders.

2.5 Acknowledgements

- Roddy Walsh: curated the list of classified variants from the DCM and HCM cohorts
- Nicola Whiffin: curated the list of variants identified in cohorts of healthy volunteers
- Rachel Buchan, William Midwinter and Alicja Wilk performed sample sequencing
- Stuart Cook and Chee Jian Pua: curated the samples of DCM cases and healthy controls from Singapore National Heart Centre.
- Magdi H. Yacoub, Yasmine Aguib and Mona Alouba: curated the samples of DCM cases and healthy conrols from Egypt Aswan Heart Centre
- Leonardo Bottolo: guidance on method design and machine learning models
- James Ware: study design; overall mentorship
- Samples from the SHaRe registry are curated by the following principal investigators: Sharlene M. Day, Euan Ashley, Steven D. Colan, Michelle Michels, Alexandre C. Pereira, Daniel Jacoby, Carolyn Y. Ho, Iacopo Olivotto, Gunnar T. Gunnarsson, John L. Jefferies, Chris Semsarian, Jodie Ingles and James Ware.

Chapter 3 Homologous residues constraint provides strong evidence to prioritise deleterious missense variants

3.1 Introduction

In Chapter 3, I develop a novel line of computational evidence to interpret missense variants and evaluate its utility.

Determining the causal relationship between genetic variants and diseases is a critical challenge in realising the promise of genome medicine. For the majority (>94%) of germline missense variants present in humans, their clinical impact remains unknown.
While recent state-of-the-art computational approaches focus on leveraging ensemble learning or deep learning methods to best summarize multiple lines of molecular evidence[60,69], strategic breakthrough also relies on developing novel and strong evidence to fully characterize variant pathogenicity.

As one of the solutions, a catalogue of natural variation found in general human populations offers a powerful resource to assess the clinical impact of variants. Since variants causing severe early-onset disorders are under strong selective pressure, they are likely to be observed less often in the general population compared with neutral variations. The degree of deviation from the number of observed variants to the number of variants expected by chance under neutral selection is quantified as a measure of genetic constraint. Measuring genetic constraint has been demonstrated to provide strong evidence to discover disease-associated genes[19,20,119], identify critical regions within genes susceptible to have deleterious variants[26,27], and investigate the effect of non-coding variants in regulatory elements[31,32,120].

While these existing metrics are useful to prioritise critical regions along with the linear space of the genome, their efficacy remains insufficient for situations in which pathogenic missense mutations could distribute sparsely within the genes or sub-genic regions. To address this issue, we sought to develop an amino-acid level constraint. Given that we expect to observe one missense variant for every six bases in exome from the current sample size in gnomAD (5,206,202 missense variants observed out of 30 Mb as the size of exome), we are still limited to evaluate depletion of variants at single residues. But instead, we could evaluate a group of residues. While existing approaches assessed linear regions along the genome, we aggregate the genetic constraint signals over homologous positions in protein domain families as they are likely of similar functional relevance. Analyses based on homologous residues across domains have been applied successfully to predict functional residues[121,122]. Furthermore, genetic intolerance of Pfam domains is also found to have low deviation across individual homologous domains[123].

Here I develop Homologous Residue Constraint (HRC), which is a novel constraint metric to evaluate the depletion of missense variants over homologous residues in proteins. In validation, I demonstrate that the variant prioritised by HRC are highly associated with known disease-causing variants. In comparison to existing gene- or region-level constraint metrics, HRC is complementary. It has especially high precision in prioritising missense variants in protein domains. We found *de novo* variants (DNVs) disrupting constrained residues are significantly enriched in both probands with neurodevelopmental disorders (n=5,264) and autism spectrum disorders (n=6,430) compared with control individuals (n=2,179). Using DNVs from 31,058 patients with developmental disorders, missense variants affecting constrained residues show excess fold-enrichment over background variation, with a similar effect size as protein-truncating variants. Finally, I demonstrate that HRC can be applied to improve gene discovery in both developmental disorders and inherited conditions. Overall, it provides an orthogonal and strong quantitative measure to prioritise deleterious missense variants.

3.2 Methods

3.2.1 Identification of homologous residues from domain family alignments

The family alignments of all 6,196 human protein domains generated using NCBI sequence database were downloaded from Pfam database[124] (the data file Pfam-A.full.ncbi.gz of release version 32.0). Given a multiple sequence alignment of a domain family, amino acids in the same column of the alignment are considered as homologous.

3.2.2 Annotation of molecular consequences of variants

RefSeq Select transcripts are used throughout the whole analysis such that each protein-coding gene has a single high-quality representative transcript. The consequences of variants are annotated by VEP (release 101)[4]. Only single-nucleotide variants with VEP annotated as “missense_variant” were included in the analysis.

3.2.3 Developing a selection-neutral, sequence-context mutational model

To estimate the number of neutral substitutions expected on a single nucleotide, I constructed a neutral mutational model using gnomAD reference population. Previous studies have shown that the mutation rate of single nucleotide substitution under neutral selection could be predicted based on sequence context and methylation level[19]. Given the baseline substitution rate using a tri-nucleotide sequence text model estimated from variants in intergenic or intronic regions by gnomAD[20], I calibrated the baseline mutation rate to probabilities of neutral substitutions within the 125,478 exomes in gnomAD following the procedures described in the gnomAD flagship paper[20].

I firstly used linear regression to predict proportions of neutral substitutions given the baseline mutation rates. For each possible tri-nucleotide sequence context, the proportion of neutral substitutions is calculated as the ratio of observed synonymous substitutions over all possible synonymous substitutions. For example, to calculate the proportion of neutral substitutions from AAT to AGT, we firstly find the number of all possible synonymous variants introduced by mutating AAT to AGT along exome and then count the ones observed in gnomAD v2 exome data. This ratio of observed to all possible numbers is used as the dependent variable in linear regression. Since the observation of substitutions would be biased by sequencing coverage, at this step only sites with high coverage (median depth 40) are included in the regression. Two linear regression models were fitted, one for substitutions at CpG sites and the other one for non-CpG sites (**Figure 3.1**). The methylation data for CpG sites was downloaded from gnomAD public datasets and was categorised into three bins: low, medium and high methylation levels as previously described[20]. There are 8 possible CpG sites considering trinucleotide context: ACG, TCG, GCG, CCG and their complementary sequences. As each CpG site is further split into three methylation levels, there are 24 possible substitutions specified by trinucleotide context and methylation level. For non-CpG sites, there are 184 possible substitutions given trinucleotide context. In total, we can evaluate 208 context and methylation-dependent substitutions. With these predicted probabilities of substitutions, we can estimate the expected number of single-nucleotide variants under neutral selection (Expected) in the 125,478 exomes in gnomAD.

Secondly, I adjusted the probabilities of neutral substitutions for low-coverage sites (median depth<40). To this end, the Observed/Expected ratios for synonymous variants were aggregated for each sequencing coverage. Given a sequencing coverage, it is calculated as : the expected number of variants is the sum of predicted proportions of neutral substitutions for each site derived from the first step, indicating the number we expect with high-coverage sequencing; the observed number of variants is the sum of observed synonymous variants for each site. A linear model is fitted to predict the Observed/Expected ratios given a sequencing coverage on a $\log_{10}$ scale (R^2=0.96, *P-value*=2.2×10^{-16}; **Figure 3.2)**. The predicted Observed/Expected ratios by the model are used as correction factors to adjust the expected number of variants at low-coverage sites.

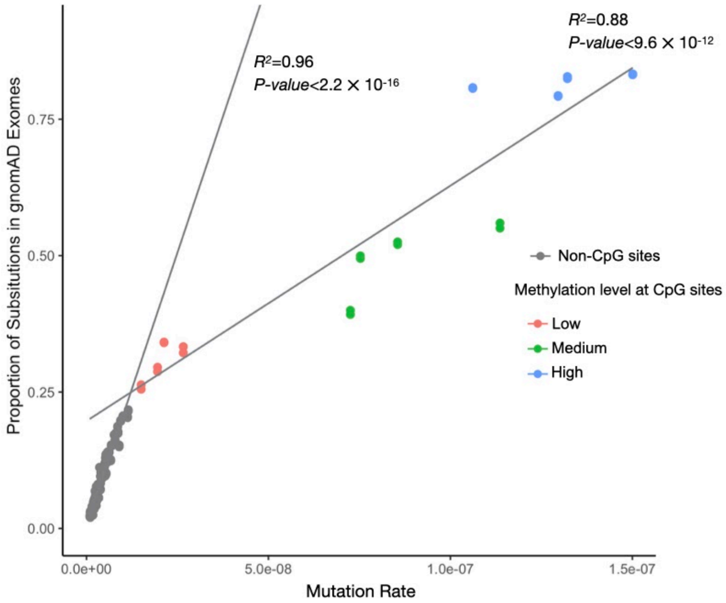


Figure 3.1 Calibration of baseline mutation rates to probabilities of neutral substitutions. Two linear regression models were fitted to predict the proportions of neutral substitutions within the 125,478 exomes from gnomAD: one for CpG sites and the other one for non-CpG sites. This shows that the model is well calibrated for the effect of CpG methylation. In the plot, each dot represents a type of substitution specified by trinucleotide sequence context and methylation level (for CpG sites).

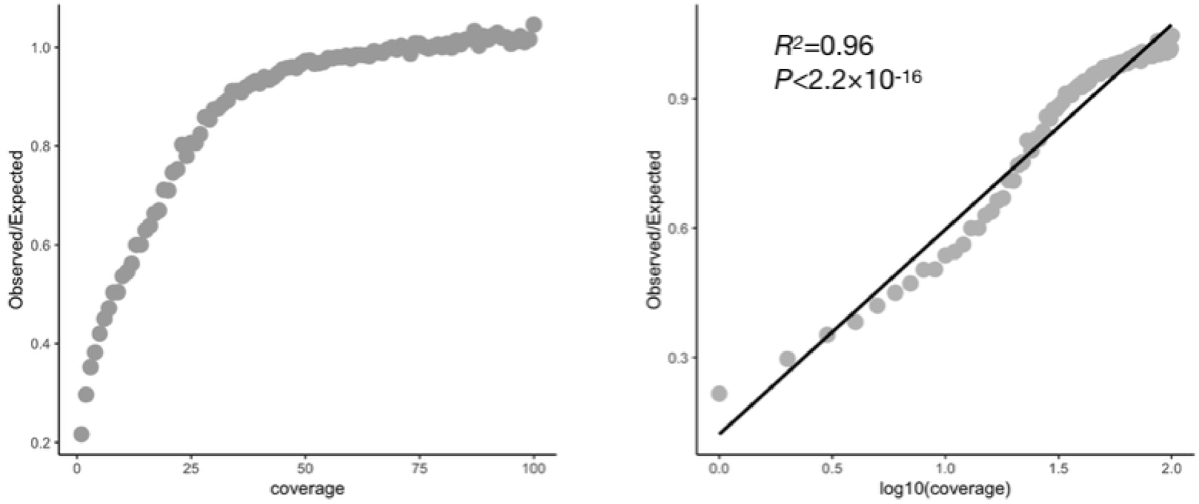


Figure 3.2 Calibration of probabilities of neutral substitutions on low-coverage sites (coverage<40). Here a linear model is fitted to predict the Observed/Expected ratios given a sequencing coverage on $\log_{10}$ scale. The predicted Observed/Expected ratios are used as correction factors to adjust the expected number of variants at low-coverage sites.

3.2.4 Estimating Homologous Residue Constraint

An overview of measuring Homologous Residue Constraint is illustrated in **Figure 3.3**.

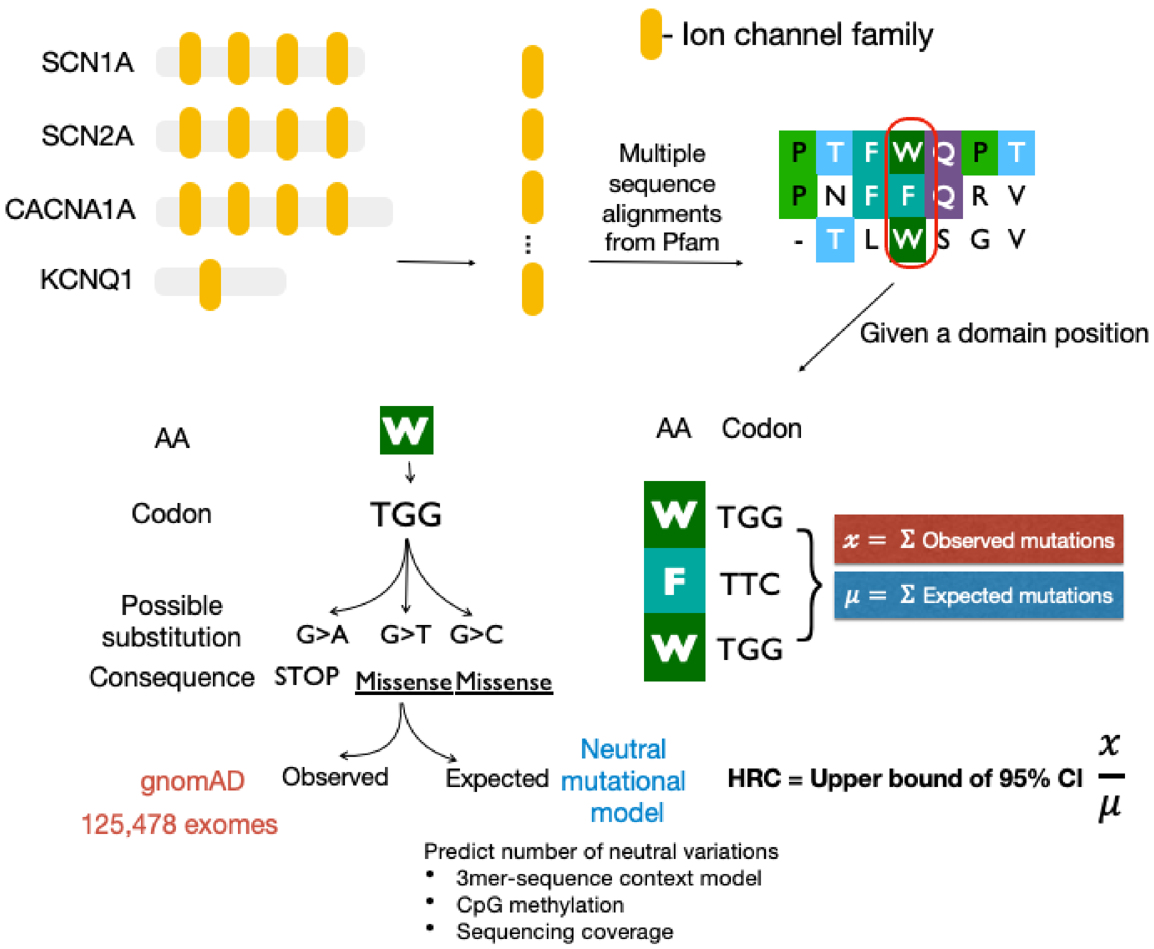


Figure 3.3 Overview of developing Homologous Residue Constraint scores.

For an aligned position in a Pfam domain family, we assessed all possible missense substitutions in residues on the position. Among all the possible missense substitutions, the number of overserved substitutions in gnomAD is counted (Observed). The expected number of missense substitutions is calculated as the sum of predicted probabilities of neutral substitutions given by the neutral mutational model (Expected). The genetic intolerance of this aligned position is calculated as the ratio of Observed/Expected.

In order to control the quality of assessing genetic constraint in homologous residues, I excluded any domain position with less than three expected variants in this analysis since it indicates that the number of possible missense variants in this position is too small to evaluate genetic constraint robustly. If the number of observed substitutions follows Poisson distribution under the null hypothesis (no selection), even with the zero observed substitution, the expected number needs to be at least three to reach significance threshold (the probability of observing zero occurrences with mean occurrence as three is 0.049. In R, it is calculated as “ppois(0,3)=0.049”). It might also indicate the corresponding column is constructed with low confidence filled with a large proportion of gaps (>95% in our observation). Filtering these columns would also limit the effect of alignment bias on defining homologous residues.

Homologous Residue Constraint is defined as the upper limit of 95% confidence interval for the Observed/Expected ratio. The confidence interval for the Observed/Expected ratio is estimated using a Bayesian approach[20]. The unknown true Observed/Expected ratio (constraint) is considered as a random variable with a uniform prior between 0 and 2. The likelihood function for a given constraint value is given as the Poisson density:

$$\Pr(X = Observed|constraint = \lambda) = \frac{(\lambda * Expected)^{Observed} e^{-\lambda * Expected}}{Observed!}$$

Thus, the posterior probability of a given constraint value could be derived by:

$$\Pr(constraint = \lambda|Observed, Expected) = \frac{\Pr(constraint=\lambda) * \Pr(X=Observed|constraint=\lambda)}{\sum_{constraint} \Pr(constraint=\lambda) * \Pr(X=Observed|constraint=\lambda)}.$$

We could further obtain the 95% confidence interval of constraint by taking the 2.5% and 97.5% quantile from its posterior probability distribution. Therefore, the upper bound of 95% CI is taken as the constraint score of homologous residues (HRC). If a residue is scored as

HRC <1, it indicates that missense variants disrupting the given domain position are significantly (*P-value* < 0.05) depleted of variants thus under selection pressure.

3.2.5 Evaluating the pathogenicity of ClinVar variants

I tested the association of HRC with known disease-causing variants by using ClinVar variants. ClinVar VCF file was downloaded from ClinVar public FTP site (version 20201114 https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/archive_2.0/2020/clinvar_20201114.vcf.gz). I extracted 22,886 Pathogenic/Likely pathogenic missense variants in Pfam domains, whose clinical significance was recorded as “Pathogenic”, “Likely_pathogenic” or “Pathogenic/Likely_pathogenic”. 7,137 Benign/Likely benign variants in Pfam domains were extracted with clinical significance recorded as “Benign”, “Likely_benign” or “Benign/Likely_benign”. After keeping the HRC assessable domain positions, 13,009 Pathogenic/Likely pathogenic and 3,914 Benign/Likely benign variants were used as test data. Only variants with no conflicting interpretation were included in the test set.

3.2.6 Evaluating the pathogenicity of *de novo* variants

To test the enrichment of DNVs prioritised by HRC in affected individuals versus unaffected individuals, I analysed the published DNVs in 5,264 patients ascertained with neurodevelopmental disorders, 6,430 patients ascertained with autism spectrum disorder, and 2,179 unaffected controls curated by Satterstrom et.al[125]. Of the missense DNVs in cases of NDD, in cases of ASD and in unaffected controls, the following number of variants is assessed by each tool (*M* DNVs in NDD; *N* DNVs in ASD; *P* DNVs in controls): HRC (891; 702; 219), CCR (2,597; 2,379;735), RMC (1,172; 894; 243) and para_zscore (1,492; 1,314;414).

I applied an independent approach to measure the accuracy of predicting damaging *de novo* missense variants by testing the enrichment of DNVs prioritised by HRC in affected individuals versus neutral variants estimated by a null sequence-context based *de novo* mutational model[24]. This measurement can also be used to assess whether HRC could distinguish pathogenic and benign variants within disease genes as the enrichment of DNVs in cases vs control individuals can be driven by gene-level disease association. I analysed the published DNVs in 31,058 patients with developmental disorders. Of the 28,193 missense DNVs in the cohort, the following number variants is assessed by each tool: HRC (6,787), CCR (16,401), RMC(8,456) and para_zscore(11,270). The burden of DNVs is calculated as the ratio of the number of observed DNVs to the number of expected DNVs. The number of observed DNVs is directly counted from the variants seen in the cohort. The number of expected DNVs under neutral selection for the cohort is calculated by summing the product of the trinucleotide *de novo* mutation rate (could be obtained downloaded at https://github.com/jamesware/denovolyzeR-ProbabilityTables/blob/master/data-raw/fordist_1KG_mutation_rate_table.txt or from Wellcome Sanger FTP at ftp://ftp.sanger.ac.uk/pub/project/ddd/rates) and the number of exome samples (2×31,058) for each nucleotide. The effective sample size for X-chromosome is adjusted considering sex-chromosome transmission as previously described[126]. Assuming the number of observed DNVs follows a Poisson distribution, the 95% confidence interval for the mean number of observed DNVs could be estimated by using an exact method. In R, it is calculated as “poisson.test(n_obs, conf.level=0.95)”.

3.2.7 Testing improving power of gene discovery

To demonstrate the utility of applying HRC to discover more disease genes reliably, I upgraded the gene-specific de novo weighted enrichment simulation test (DeNovoWEST)[126] by adding HRC to score missense variants. In the original framework of DeNovoWEST, the weight of a missense variant used in the simulation test depends on the regional missense constraint. Here I incorporated HRC into this framework through the following procedures: (1) combining HRC with regional missense constraint to label constrained missense variants, thus a missense variant is considered as constrained if either RMC or HRC score it as constrained; (2) updating the weights of missense variants used in DeNovoWEST: I calculated the burden of *de novo* missense variants against a null *de novo* mutational model[24] and inferred the corresponding positive predictive values (PPV) for all possible categories using constraint (based on step 1) and CADD scores. The newly derived PPV is used as weights in the downstream gene-specific test. The upgraded test was applied in the full cohort of 31,058 parent-proband trios of developmental disorders[126].

3.3 Results

3.3.1 Measuring homologous residues depletion of missense variants across human domains

A graphic illustration of the method is shown in **Figure 3.3**. 70 million all possible rare (defined as gnomAD MAF<0.1 or unobserved) missense variants in 19,212 human genes (annotated with RefSeq select transcripts) were mapped to protein domain families based on Pfam database[124]. There are 28,032,394 all possible rare missense variants in 15,305 genes (out of 19,212 genes) composed of 5,807 Pfam domains. After excluding domain positions with poor statistical power (see **3.2.4**), there are 15,236,101 possible rare missense variants from 699 Pfam families with 78,070 domain positions assessable in 9.918 genes. To be noticed, though we only assess 12% Pfam domains after quality control, disproportionately more than 50% possible missense variants occur in these 699 domains.

To quantify the genetic intolerance/constraint of missense variants of a homologous domain position, I calculated the Observed/Expected ratio for all the residues in the position, which is the ratio of the number of rare missense variants (Observed) observed in the gnomAD reference population (v2.1.1 exome datasets 125,748 individuals) to the number of neutral substitutions expected (Expected) to occur in a reference population with a sample size of gnomAD. The number of neutral substitutions is the sum of predicted mutability given by a neutral mutational model taking account of tri-nucleotide sequence context, CpG methylation levels, and sequencing coverage (see **3.2.3)**. The upper bound of 95% CI of the Observed/Expected ratio is derived and defined as the Homologous Residue Constraint (HRC) (see **3.2.4**). A protein residue with HRC score smaller than 1 indicates that missense variants affecting homologous residues are significantly under negative selection and likely to be deleterious.

There are 3,304,332 possible missense variants in 9,085 constrained positions from 596 Pfam domains (HRC<1; 21.7% of assessable variants) and 1,322,835 possible missense variants in 3,381 highly constrained positions in 458 Pfam domains (HRC<0.8, a threshold we find clinically relevant in various applications demonstrated below).

3.3.2 Homologous residues constraint improves precision to detect pathogenic variants

Under the assumption that genetic intolerance is predictive of clinical importance, I would expect an enrichment of pathogenic variants at constrained domain positions. Therefore, I tested whether constrained domain positions are enriched with known disease-causing variants compared to benign variants. Of all missense variants from domain families in ClinVar (22,886 Pathogenic/Likely pathogenic and 7,137 Benign/Likely benign variants), we were able to analyse a total of 13,009 Pathogenic/Likely pathogenic variants and 3,914 Benign/Likely benign variants. I found that ClinVar pathogenic variants are significantly enriched at constrained domain positions (HRC<1: OR=6.1, 95%CI=5.5-6.8) and significantly depleted at unconstrained domain positions (HRC>=1: OR=0.16, 95%CI=0.14-0.18) (**Figure 3.4a**). The association increases as domain positions are under stronger genetic constraint indicating that variants disrupting these positions are more likely to cause diseases.

To characterise HRC performance to prioritise variants, I compared its ability to rank pathogenic variants with the existing sub-genic constraint models: Regional Missense Constraint[26] (RMC) and Constraint Coding Region[27] (CCR), and a sequence conservation metric of human paralogous residues measured by para_zscore[17]. As each approach focuses on generating predictions on different areas of the exome given different hypotheses, I analysed a consensus set of ClinVar variants that can be prioritised by all four methods (3,661 pathogenic variants and 537 benign variants). Among all the benchmarked scores, CCR has the highest area under the Precision-Recall Curve and HRC comes as the second best (CCR: 98.0%; HRC: 96.7%; RMC=94.2%; para_zscore: 94.2%) (**Figure 3.4b**).

Importantly, I examine the capability of each method to precisely identify pathogenic variants measured by diagnostic odds ratio and precision (Positive Predictive Value) especially in top-ranked variants. As in a false-positive intolerant setting such as clinical genetic diagnosis, top-ranked variants (highly prioritised) are more likely to receive greater attention for downstream validation or help to inform clinical decision making. Compared with existing constraint or homologous residue-based metrics, HRC has a higher odds ratio to discriminate pathogenic variants from benign variants in the highly prioritised variants up to the top 40% (approximately HRC <0.8) shown in **Figure 3.4c**. Correspondingly, it indicates that HRC model has a particularly higher precision of predicting pathogenic variants given the same true positive rate among the highly prioritised variants (**Figure 3.4b**). To be noticed, HRC outperforms para_zscore, which suggests that genetic constraint derived from natural variation in human populations is highly relevant of disease impact compared to sequence conservation.

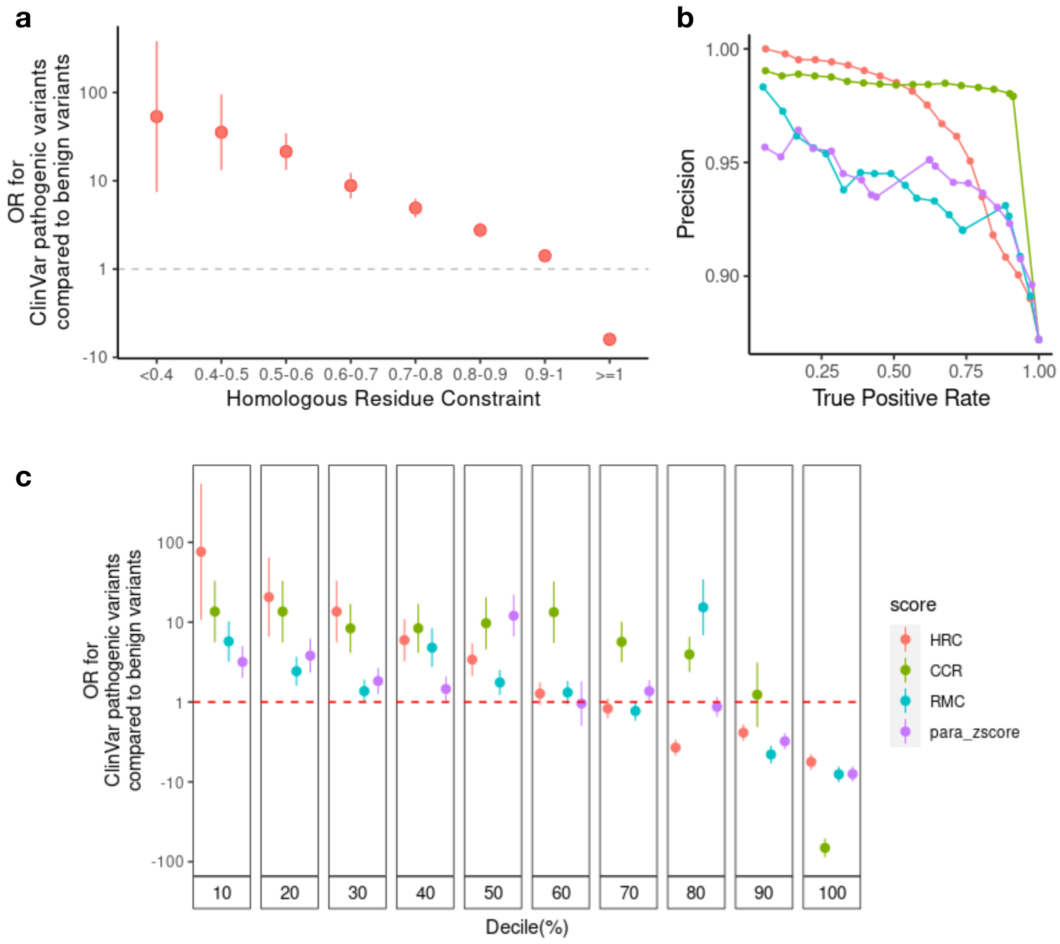


Figure 3.4 Pathogenic variants are significantly enriched in constrained homologous residues in domain families compared with benign variants. (a) Odds Ratio (OR) measures the association between homologous residue constraint and pathogenicity. **(b-c)** HRC improves the precision of discriminating pathogenic from benign variants in protein domains. In highly prioritised variants, HRC shows a higher odds ratio and precision compared with benchmarked constraint (CCR and RMC) or homologous residue-based metrics (para_zscore) among about the top 40% ranked variants. **(b)** The Precision-Recall curve demonstrates that HRC has higher precision over the other methods in top-ranked variants. **(c)** Odds Ratio measuring the enrichments of pathogenic variants versus benign variants in each decile.

3.3.3 HRC identifies highly deleterious *de novo* missense variants

Across the full spectrum of pathogenicity of germline variants, a subset of *de novo* variants (DNV) could reduce reproductive fitness the most (pathogenic) and most often (penetrant) since they haven't been transmitted to one generation. As every exome carries one DNV on average (a genome-wide rate at 1.2×10^{-8} mutations per nucleotide per generation[127]), we would expect DNVs from cases are enriched with highly constrained variants compared with controls. Therefore, we want to ask whether highly constrained residues identified by HRC could provide a novel line of evidence to interpret *de novo* variants causing severe developmental disorders.

I first examine the association between *de novo* variants disrupting constrained domain positions and diseases. By analysing published *de novo* variants identified in 5,264 probands ascertained with severe neurodevelopmental delay (NDD) and 2,179 unaffected individuals[125], I find *de novo* missense variants in highly constrained domain positions (HRC<0.8 or about the top 20% percentile) are significantly enriched in cases (OR=5.2, 95% CI=3.0-9.1), which is comparable with applying the existing constraint or homologous-residue based metrics (**Figure 3.5a-c**). Similarly, I also find highly constrained *de novo* missense variants are significant enriched in cases ascertained with autism spectrum disorders (ASD) though with weakened effect size (OR=2.4, 95% CI=1.3-4.3; **Figure 3.5d**), which is likely due to the differences in genetic architecture. Variants observed in gnomAD are excluded here as they are unlikely to cause developmental diseases with high penetrance[128].

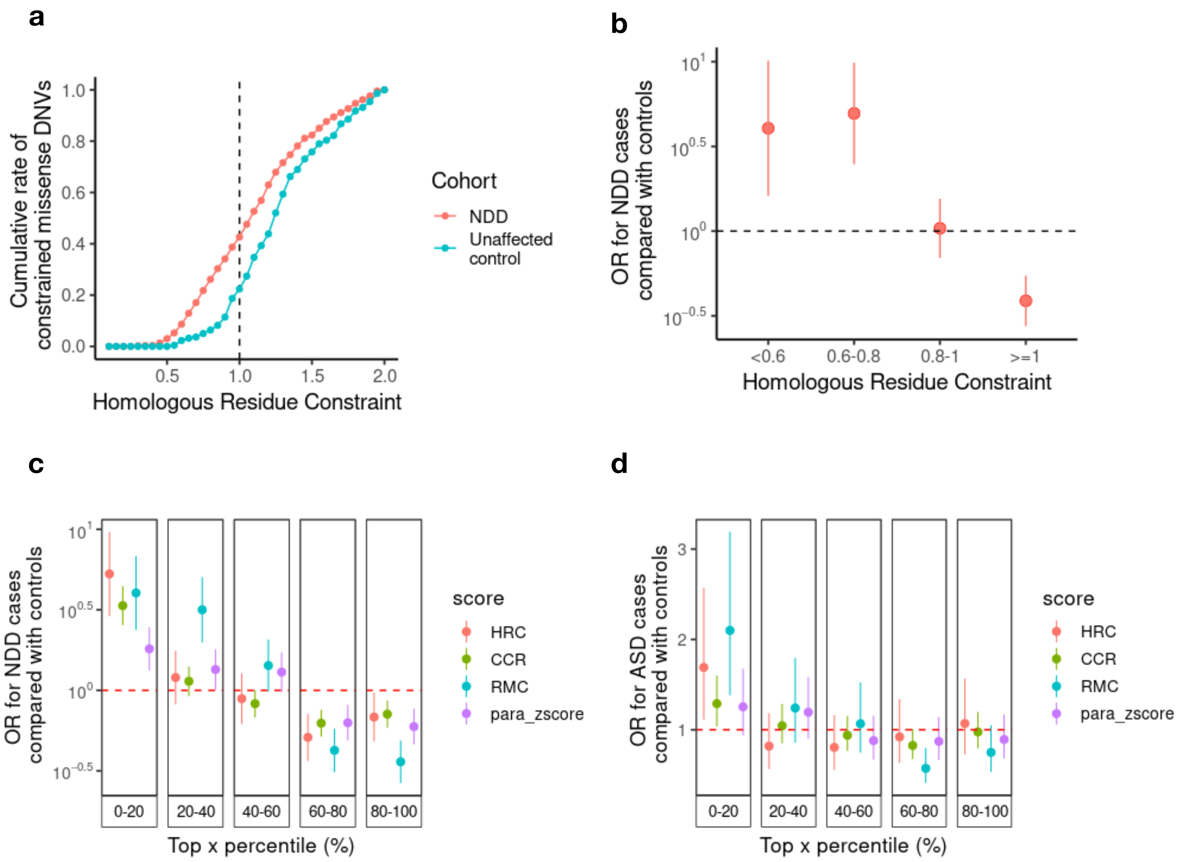


Figure 3.5 *De novo* missense variants affecting highly constrained domain positions are enriched in cases in neurodevelopmental developmental disorders versus unaffected controls. (a-c) Enrichment of constrained missense DNVs in 5,264 NDD-ascertained cohort versus 2,179 unaffected controls. **(d)** Enrichment of constrained missense DNVs in 6,430 ASD-ascertained cohort versus 2,179 unaffected controls.

Since the above enrichment of damaging *de novo* variants in cases could be largely driven by gene-level disease association, I sought to dissect whether pathogenic *de novo* variants could be discriminated from benign *de novo* variants in disease genes. I compared the burden of *de novo* variants in a larger cohort with 31,058 patients of developmental disorders[126] (DD) against background expectation from a null *de novo* mutational model[129]. Applying HRC to prioritise damaging missense variants could increase the burden of *de novo* missense variants from baseline 1.23-fold (95%CI=1.22-1.25) to 3.34-fold (HRC<0.8, 95%CI = 3.16-3.53 ; for the burdens across ranges of constraint values see **Figure 3.6a**), even significantly higher than that of protein-truncating variants (PTV; 2.32-fold, 95%CI=2.24-2.39) (**Figure 3.6a**). Compared with other approaches, the top 5% *de novo* missense variants prioritised by HRC show the highest burden of 3.9-fold (95%CI=3.7-4.2), indicating a set of *de novo* missense variants occurring in cases nearly four times more than expectation (**Figure 3.6b**). When we specifically restrict our analysis on 285 previously identified DD-associated genes[126], HRC clearly outperforms other approaches to distinguish pathogenic variants and background benign variants even in the same set of disease genes consistently across the spectrum of percentiles (**Figure 3.6c-d**). The burden of damaging *de novo* missense variants in cases is 32-fold (95%CI: 29-35) higher than background variants in the top 5% of most constrained residues with a comparable effect size of protein-truncating variants (32-fold, 95%CI: 30-34), highlighting that HRC is highly precise to prioritise pathogenic *de novo* missense variants in disease genes. I also found that para_zscore also outperformed the regional constraint scores, indicating that homologous residues-based approaches could be more predictive of pathogenicity within disease genes.

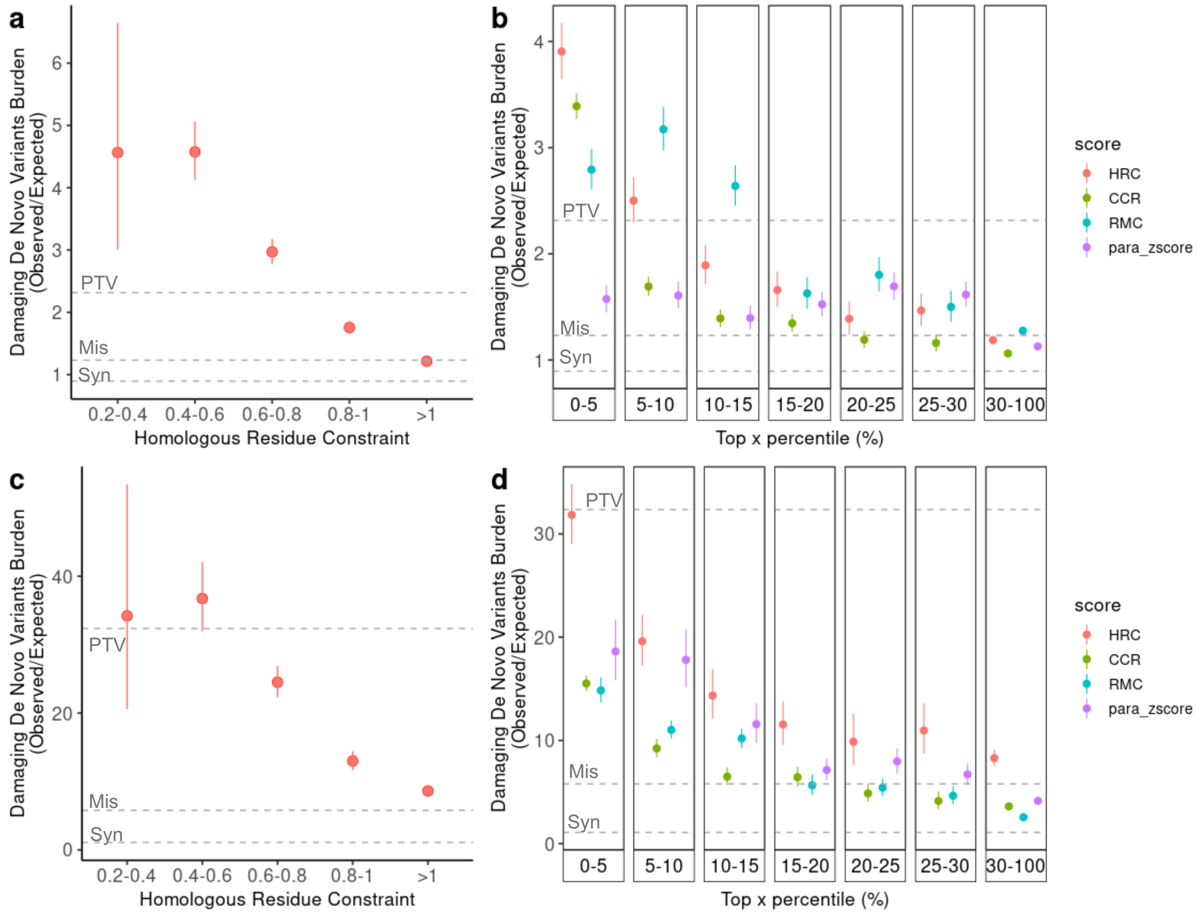


Figure 3.6 HRC prioritises damaging *de novo* variants with significantly higher burden compared with background variants in 31,058 parent-proband trios of developmental disorders. (**a-b**) Burden of damaging *de novo* missense variants in all genes. (**c-d**) Burden of damaging *de novo* missense variants in DD genes.

3.3.4 HRC can also improve gene discovery

Given the precision of HRC in prioritising pathogenic missense variants, I further investigate whether it could be applied to improve gene discovery. I updated a gene-specific *de novo* weighted enrichment simulation test (DeNovoWEST)[126] to incorporate HRC score to weight missense variants (see **3.2.7**). With the upgraded test, there are 286 genes identified in the full cohort of 31K DD trios and 97 identified in the undiagnosed cohort of 24K DD trios (probands who do not carry pathogenic variants in consensus diagnostic genes, as previously defined[126]) reaching the genome-wide significance threshold (multiple testing *P*-value<0.05/(2*18,762) using Bonferroni correction, taking account of testing in 18,762 genes and 2 tests per gene). There are seven candidate novel genes identified across the two tests, which were not considered as DD genes (not reach genome-wide significance threshold) in the original study[126]. These novel candidate DD genes are likely to act through an altered-function mechanism as opposed to a loss-of-function mechanism since they have more constrained *de novo* missense variants by HRC than *de novo* PTVs observed in the patient cohort (**Figure 3.7**). This highlights that HRC could be useful in discovering genes with altered-function mechanisms, overcoming the limits of current approaches.

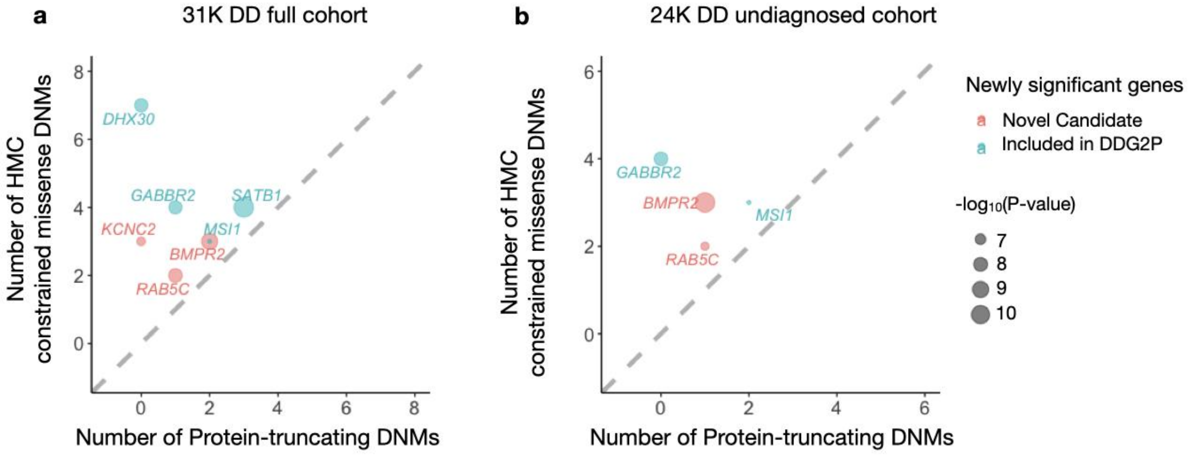


Figure 3.7 *De novo* variants identified in 31,058 parent-proband trios reveal seven genes associated with developmental disorders at genome-wide significance for the first time in the full DD cohort (a) and the previously-undiagnosed subset (b). Four of these genes have been previously curated as DD genes on the basis of other lines of evidence, and are already included in the G2P database as established Developmental Disorder genes (blue), while three genes represent new candidate DD genes (red). Numbers of constrained missense DNMs classified by HRC and protein-truncating DNMs were compared. The newly-significant associated genes likely act through altered function mechanisms as there are more constrained missense variants than PTVs.

I also performed a case-control gene burden test in 6,327 patients of hypertrophic cardiomyopathy (HCM) from SHaRe registry[57] and gnomAD general population of 125,748 individuals, which are curated in the study of Chapter 2. We focus on eight sarcomere genes here, which are fully genotyped in the patient cohort. Four sarcomere genes have at least one missense variant in protein domains carried by the patients (*MYBPC3, MYH7, ACTC1, MYL2*). Collapsing highly constrained (HRC<0.8) or nominally constrained (HRC<1) missense variants can increase the gene-disease ORs compared with unconstrained or unclassified missense variants, which demonstrates the utility of HRC in adult-onset disorders (**Figure 3.8**). Burden test in *MYBPC3* shows significantly elevated associations in both highly constrained versus constrained variants, and (highly) constrained versus unconstrained/unclassified variants. We expect collapsing analysis using HRC would have better statistical power in genes with multiple domains, which is exemplified by *MYBPC3* here (with 12 domains while the other seven sarcomere genes have 0-3 domains).

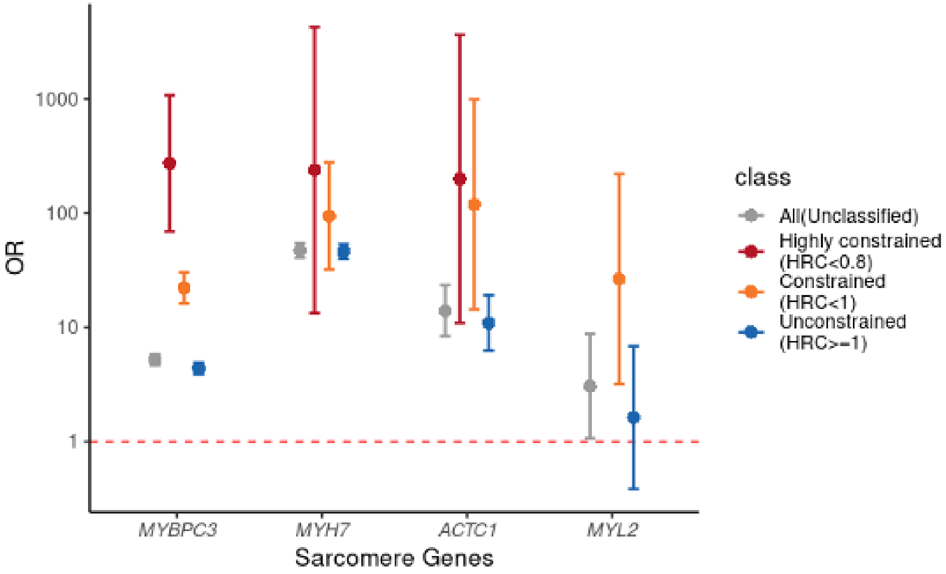


Figure 3.8 Applying HRC in gene-burden tests of rare missense variants in HCM cases and controls. Four sarcomere genes with missense variants in Pfam domains from affected

individuals were assessed. Highly constrained (HRC<0.8) or nominally constrained missense variants (HRC<1) have increased association with HCM compared with controls.

3.3.5 HRC is orthogonal to existing constraint metrics to prioritise missense variants

One characteristic about HRC is that it can estimate genetic constraint at the amino-acid level. To demonstrate this, I compared the relationship between HRC scores and the existing constraint metrics including genic constraint scores (LOEUF and MOEUF, measuring genic intolerance of loss-of-function variants and missense variants respectively), regional constraint scores (CCR and RMC, measuring the genetic constraint of sub-genic regions). As each method has its own strengths and weaknesses, I focus on whether HRC could complement existing metrics in variant prioritisation.

I found constrained homologous residues are distributed across full ranges of these existing metrics in either constrained genes/regions or unconstrained genes/regions. Concordantly, if a gene/region is more constrained as a whole, on average it also has more constrained homologous residues compared to a less constrained gene/region (**Figure 3.9**). What's more important, for genes/regions considered as unconstrained, there are substantial numbers of highly constrained missense variants that could be prioritised by HRC (HRC<0.8): 254,226 in genes nominally unconstrained of loss-of-function variants (LOEUF>=1), 342,065 in genes nominally unconstrained of missense variants (MOEUF>=1), 2,772,371 in CCR unprioritized regions (<95th percentile) and 988,686 in RMC unprioritized regions (>0.8 or unscored).

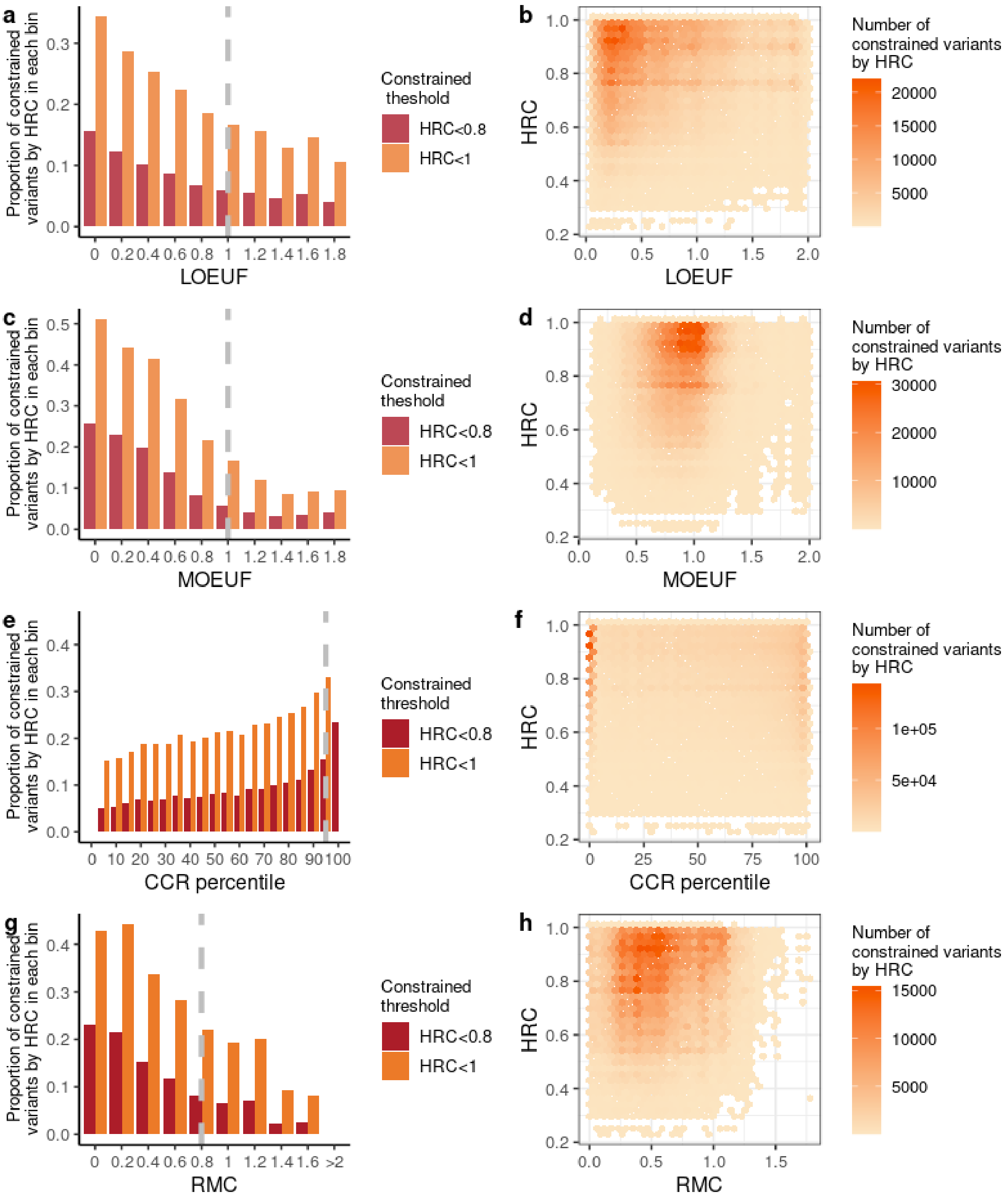


Figure 3.9 Comparing the distributions of HRC score and other gene-level and regional level constraint scores. Bar plots in the first column display the proportion of constrained missense variants by HRC across genes or regions. 2D-bin plots in the second column display the counts of constrained missense variants given HRC score and gene/region score. (a-b) The relationship between HRC and a gene's LOEUF score (genetic constraint of loss-of-function variants; a lower value indicates higher constraint). A gene with

LOEUF<1 (grey dashed line) is under nominally constrained (c-d) The relationship between HRC and a gene's MOEUF score (genetic constraint of missense variants; a lower value indicates higher constraint). A gene with MOEUF<1 (grey dashed line) is nominally constrained. (e-f) The relationship between HRC and CCR (a higher percentile indicates higher constraint). A region with CCR percentile >95% (grey dashed line) is considered as constrained recommended by authors[27]. (g-h) The relationship between HRC and RMC (a lower value indicates higher constraint). A region with RMC>0.8 (grey dashed line) is considered unconstrained.

3.4 Discussion

Mining the pattern of natural variations depleted in the human population uncovers variants of strong clinical impact. Here I describe a novel form of genetic constraint signal, Homologous Residue Constraint to interpret missense variants. Compared with existing metrics measuring genetic constraint over linear space of the genome, HRC considers the "vertical" space of the genome, which enables us to assess genetic constraint at the amino-acid level. Compared with existing constraint or homologous residue-based score, I found HRC is highly precise to predict pathogenic variants in ClinVar.

An important application of HRC would be to predict the deleteriousness of novel missense variants. Applying HRC to identify *de novo* variants affecting constrained positions, I found patients with severe developmental delay including neurodevelopmental disorders and autism spectrum disorders carry highly constrained missense *de novo* variants significantly more often than unaffected controls and background *de novo* variants. HRC also has high precision to discriminates pathogenic *de novo* variants from benign *de novo* variants in known DD genes, with effect size comparable to protein-truncating variants.

As the statistical power for existing genetic constraint metrics depends on coding sequence length either implicitly (LOEUF, MOEUF, and RMC) or explicitly (CCR uses length as a covariate to measure genetic constraint) (**Figure 3.10**), HRC provides an approach to prioritise variants independent of gene/region length but depends on repeated domains in human proteins. Thus, it could also help to identify novel disease genes, which is demonstrated by incorporating HRC into the gene-discovery framework DeNovoWEST to identify nine additional novel candidate DD genes likely acting through an altered-function mechanism. The utility of HRC for this purpose is also highlighted by an intensive study to predict deleterious *TTN* missense variants in Chapter 4.

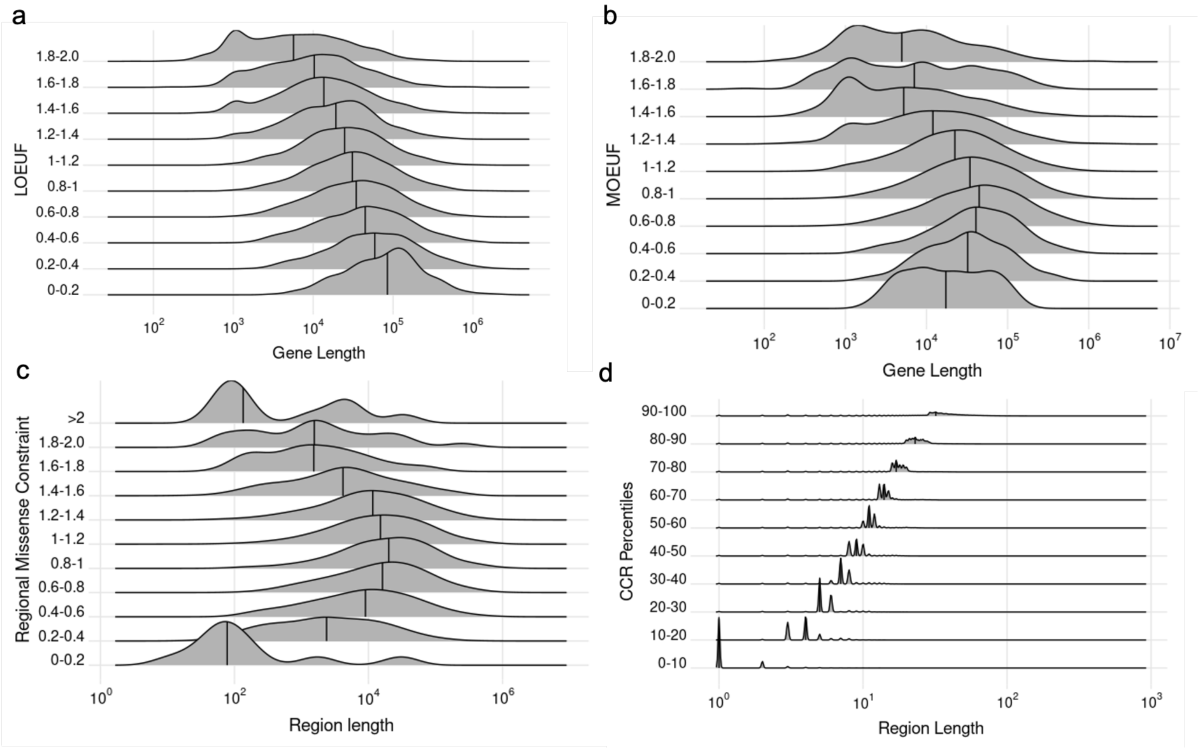


Figure 3.10 Comparing the distributions of gene or sub-genic region length across gene-level and regional level constraint scores. In specific, for all genes or sub-genic regions with a benchmarked score in a certain range, the distribution of their coding sequence length and the median (indicated as the vertical line) is shown.

There are several ways to improve the development of HRC scores. Since there is no gold-standard definition for homologous amino acids, our choice is largely limited by the availability of data. In this study, I used protein domain alignment to define homologous residues because it has better coverage in exome compared with paralogous alignment and structural alignment. As the performance of HRC could be affected by the bias of sequence alignment, I also explored whether taking account of the genetic constraint of surrounding amino acids could improve the performance since the true homologous residues are likely in neighbouring columns if not aligned with each other. This experiment shows that adding more surrounding amino acids could improve sensitivity but also compromise precision (positive predictive value) since there could be more non-relevant residues added to dilute the signal (**Figure 3.11**). To favour precision over sensitivity, I did not consider adding surrounding amino acids in our final metric and used the vanilla version.

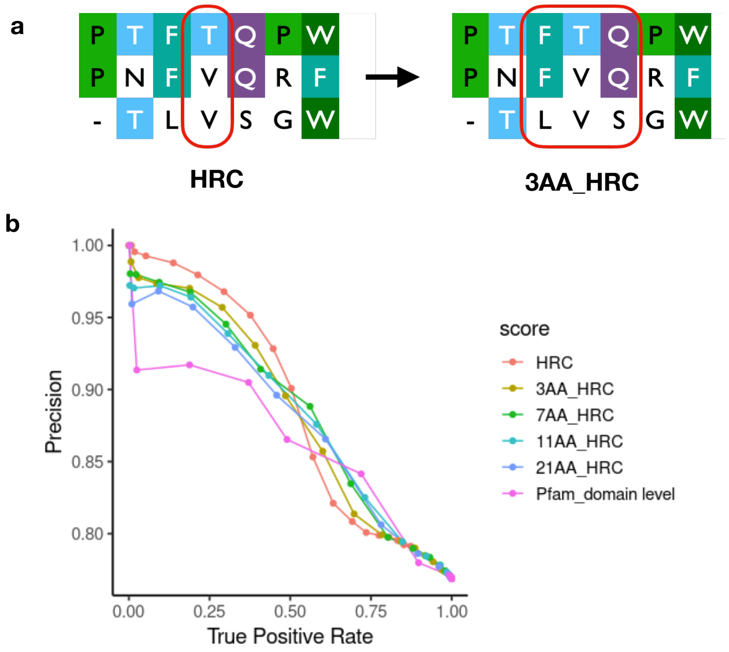


Figure 3.11 Exploring different genetic constraint measured in Pfam domains. The following metrics are calculated: HRC (vanilla), 3AA_HRC (Constraint of homologous residues within a sliding window of 3 amino acids, illustrated in **a**), 7AA_HRC, 11AA_HRC, 21AA_HRC and genetic constraint of domain-level. Comparison of their performance using Precision-Recall curves are shown in **b**.

Additional features of residues could be added to improve the positive predictive value of HRC, such as interspecies conservation and biochemical properties of aligned amino acids. As homologous residues based on sequence might not be always functionally homologous to each other, the performance of HRC could be also affected by exceptions when certain individual residues might have different functional consequences/specifications than homologous residues in their family. Though I chose to keep HRC orthogonal here without adding existing molecular evidence, there is potential for improvement by combining HRC with additional features of residues.

Like existing metrics, our ability to measure genetic constraint using gnomAD could also be affected by its inclusion of damaging *de novo* variants or recently evolved deleterious variants but not yet removed in the human population. However, as the validation in DD cases shows, the noise is neglectable compared to the true signal.

As we aggregate domains over proteins, this could increase power for some genes while decreases power for other genes. For late-onset disease genes, they might have improved predictive power by having the same domains with early-onset disease genes.

Overall, HRC provides strong evidence to prioritise missense variants. With the development of computational structural genomics in recent years, approaches like HRC could achieve even better precision using structurally aligned residues once we could obtain accurate structural models for most of the human proteins. Furthermore, the exome coverage and statistical power of HRC would be also scaled up along with the ongoing growth of large-scale population genomics data and efforts of protein family classification. Thus, HRC is a promising approach to expand our ability to interpret missense variants.

3.5 Outline of further work

- In the study, I only compared HRC with existing tools developed with similar approaches based on either genetic constraint or homologous residues. To fully characterise its performance, I am going to compare HRC with the state-of-the-art variant pathogenicity tools based on machine learning. I am also going to explore how to combine HRC with the existing computational line of evidence to best support variant interpretation.
- We will report the filtered candidate novel genes to the Developmental Disorder Genotype – Phenotype Database (also known as DDG2P)[130] to be included in the panel of DD genes.
- Compare HRC with variant effect measured through high-throughput assays. Published datasets from multiplexed assays of variant effect could be retrieved through MaveDB[131].

3.6 Acknowledgements

- Kaitlin Samocha and Matt Hurles: critical feedback and advice on future work
- Nicola Whiffin: study design; overall mentorship
- James Ware: study design; overall mentorship
- Samples from the SHaRe registry are curated by the following principal investigators: Sharlene M. Day, Euan Ashley, Steven D. Colan, Michelle Michels, Alexandre C. Pereira, Daniel Jacoby, Carolyn Y. Ho, Iacopo Olivotto, Gunnar T. Gunnarsson, John L. Jefferies, Chris Semsarian, Jodie Ingles and James Ware.

Chapter 4 The role of *TTN* missense variants in DCM

4.1 Introduction

After showing homologous residue constraint could improve variant prioritisation in Chapter 3, I applied this framework to prioritise *TTN* missense variants and study their role on DCM.

The role of TTN-truncating variants on DCM

The largest human protein titin (34,350 amino acids for its longest isoform) is a crucial component of all striated muscle. *TTN* variants are associated with both cardiomyopathy and skeletal muscle myopathies. TTN-truncating variants (TTNtv) are the most common (~15%) genetic cause of dilated cardiomyopathy (DCM)[47]. It is suggested that TTNtvs cause DCM dominantly through a loss-of-function mechanism[132]. Alternatively, it is also hypothesized that *TTN* transcripts with protein-truncating variants might be translated as toxic peptides (dominant-negative) to cause DCM[47].

TTNtv is interpretable in disease cohorts. Given a TTNtv found in patients with DCM, we can primarily distinguish the benign and likely disease-causing ones based on whether the variant is in an exon constitutively expressed in cardiac tissues. For a TTNtv in a constitutive exon of cardiac tissues, it is estimated with a probability of 0.97 to cause DCM and considered as clinically actionable[47].

In a general population, TTNtv has a frequency of ~1%, higher than the prevalence of DCM (0.4%). High-resolution cardiac imaging studies suggest that TTNtv could cause abnormal cardiac remodelling in healthy individuals though the phenotype may not reach the diagnosis criteria of DCM[132].

TTN missense variants

However, the interpretation of *TTN* missense variants is a profound challenge. There are about 235K possible *TTN* rare (gnomAD minor allele frequency < 0.1% or never observed in gnomAD) missense variants. Here I summarise the existing research attempting to evaluate the significance of *TTN* missense variants.

The first DCM-causing *TTN* missense variants were found in a family linkage study. In the first report confirming that titin variants cause DCM, it includes two families[133]. One carries a TTNtv and the other carries a *TTN* missense variant. This missense variant, Trp976Arg is predicted to disrupt a characteristic hydrophobic core structure of an immunoglobulin I-set domain by replacing the most conserved residue[134], tryptophan with a polar residue, Arginine. The disrupted immunoglobulin domain is located in the Z-disc–I-band transition zone of sarcomeres. Linkage analyses strongly support that this variant co-segregates with the disease status in the family. A later functional study further demonstrates that mutant titin protein with Trp976Arg reduces sarcomere contractile function to cause DCM in human-induced pluripotent stem cell-derived cardiomyocytes[135]. Evidently, some *TTN* missense variants are able to cause DCM.

If a variant is disease-causing, it shall be more likely to be present in patients than in healthy controls. However, rare *TTN* missense variants are collectively common in the general population[136]. As an independent analysis, I compare the prevalence of *TTN* missense variants in two cohorts including 972 patients of DCM and 676 healthy volunteer controls recruited at Royal Brompton Hospital. Both cohorts are sampled from Caucasian populations. In each cohort, about 45% of individuals carry a rare *TTN* missense variant, which is not distinguishable between cases and controls (**Figure 4.1**).

Previous studies assessing the relevance of deleterious *TTN* missense variants based on *in silico* tools (i.e. SIFT, PolyPhen2, and MutationTaster) have been inconclusive when comparing either patients with DCM versus healthy control population[136] or carriers versus non-carriers in DCM patients[137]. Recently, Herrero-Galan *et.al* suggested that missense variants that alter conserved cysteines in immunoglobin domains could have a role in DCM through modulating the mechanical properties of titin with *in vitro* evidence[138]. One missense variant that affects conserved residue Cys3892 was reported to be associated with DCM supported by family segregation data. However, the degree of contribution of *TTN* missense variants to DCM remains unknown due to a lack of population-based case-control evidence. The majority of *TTN* missense variants still largely remain uninterpreted.

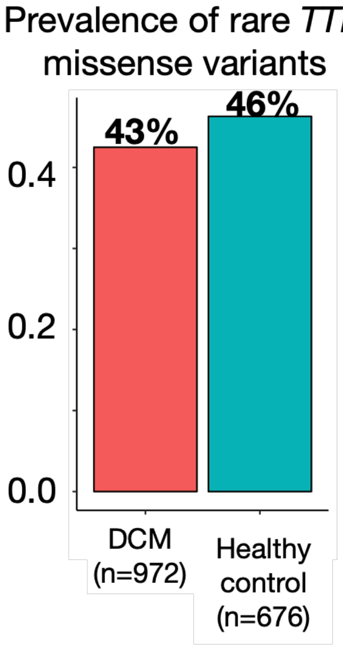


Figure 4.1 Prevalence of rare *TTN* missense variants in patients of DCM and healthy controls from Caucasians.

Here instead of assessing sequence conservation, I seek to identify genetically constrained amino acids in *TTN* and evaluate their association with DCM in population-based case-control studies. Chapter 3 describes one approach to identify constrained missense variants by aggregating the genetic constraint signal over homologous residues in protein domains. This is an approach particularly suitable to identify constrained residues in TTN. Even though TTN is a giant protein, ~70% of the titin sequence is composed of ~300 Fibronectin type-III (FN3) and Immunoglobin I-set (I-set) domains. Since there are multiple copies for the same domain, homologous residues in the same domain are likely to have similar functional relevance. I hypothesise that whether a missense variant in these domains is disease-causing, depends on the position within the domain family instead of the absolute location in the protein sequence.

In this Chapter, I am going to assess whether *TTN* constrained residues in I-set and FN3 domains are associated with DCM. Apart from using residues prioritised by HRC that are constrained across all human proteins (described in Chapter 3), I also identified TTN-specific significantly constrained residues by only considering homologous residues in TTN instead of all human proteins. To study the implication of candidate variants in human diseases, I tested the enrichment of reported disease-causing variants from ClinVar compared with background variations. To study the association between candidate variants with DCM, I also compared the enrichment of constrained missense variants in 2,023 cases of DCM compared with 2,313 healthy controls using burden tests.

4.2 Methods

4.2.1 Identification of homologous residues from domain family alignments

The family alignments of I-set and FN3 domain generated using NCBI sequence database were downloaded from Pfam database[124] (the data file Pfam-A.full.ncbi.gz of release version 32.0) A python script (*https://github.com/XiaoleiZ/parse_pfam_stockholm*) is written to extract alignment sequences from a query Pfam domain from the downloaded file.

4.2.2 Annotation of molecular consequences of variants

Consistently, *TTN* meta-transcript ENST00000589042.5 (NM_001267550.2) is used throughout the whole analysis to annotate the consequence of variants. Only rare variants (gnomAD MAF< 0.1% or unobserved in gnomAD) are considered in the evaluation. As some variants could have multiple consequences on the transcript, only the worst molecular consequence is considered. Variants were predicted to be truncating if their worst consequence included nonsense or they disrupted a canonical splice donor/acceptor sequence. The corresponding VEP annotated consequences for truncating variants include: “stop_gained”, “splice_donor_variant”, “splice_acceptor_variant” and “frameshift_variant”. In the validation of neutral mutational models on TTNtv, only SNVs are considered.

Since some missense variants could act through cryptic splicing to disrupt protein function, I also used SpliceAI to predict likely cryptic-splicing *TTN* variants. Of all possible *TTN* rare missense variants in I-set and FN3 domains, there are 528 predicted splicing variants (with recommended threshold > 0.5). They were excluded in the calculation of missense constraints.

4.2.3 Statistical testing for mutational burden

The p-value for the number of observed mutations (Observed) compared to the number of expected (Expected) (i.e. probability of number of observed mutations no more than number of expected mutations) is calculated using R as: ppois(Observed, lambda = Expected, lower.tail = TRUE).

4.3 Results

A graphic illustration of measuring TTN-specific homologous constrained residues is shown in **Figure 4.2**.

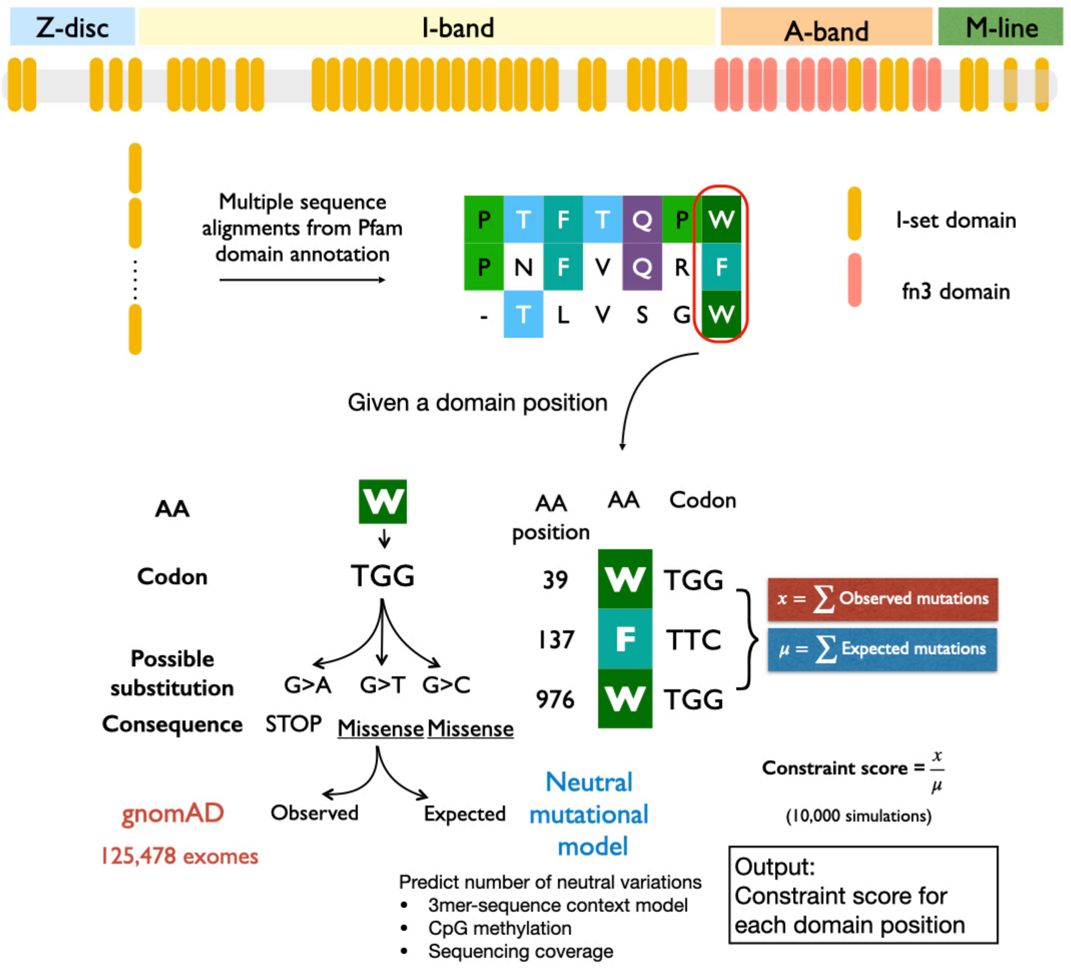


Figure 4.2 The workflow of generating homologous residue constraint in domains of *TTN*. (a) Identification of homologous residues in FN3 domains and Ig-set domains of *TTN*. (b) Calculation of observed/expected ratio for homologous residues. The number of observed substitutions is calculated from gnomAD v2.1. sequencing data of 125,478 exomes. The number of expected substitutions is estimated by a neutral mutational model taking account of 3mer-sequence context, CpG methylation level and sequencing coverage of gnomAD dataset.

4.3.1 Constructed mutational model predicts the nucleotide substitution patterns under neutrality in *TTN*

Chapter 3 describes building a neutral mutational model to predict the probability of substitutions in a population (see **3.2.3**). To validate the mutational model in *TTN*, I first applied it on the following sets of variants to evaluate their genetic constraint: *TTN* synonymous variants, and single nucleotide TTN-truncating variants. *TTN* synonymous variants are collectively and predominantly assumed to be under minimal selection, while heterozygous TTN-truncating variants are known to cause DCM. For *TTN* synonymous variants, the expected number of variants is highly correlated with the observed one from the gnomAD reference population (*TTN* synonymous variants: observed/expected=1.01, correlation coefficient *r*=0.99; **Figure 4.3a,c**). As expected, single nucleotide TTNtv occurs less often than what mutability would expect within human populations (observed/expected=0.36). Having confirmed the accuracy of the neutral mutational model, we could use it to predict the number of neutral substitutions for any coding bases in *TTN*. For *TTN* missense variants as a whole, they are not under constraint (*TTN* missense variants: observed/expected=1.02, correlation coefficient *r*=0.97; **Figure 4.3b**).

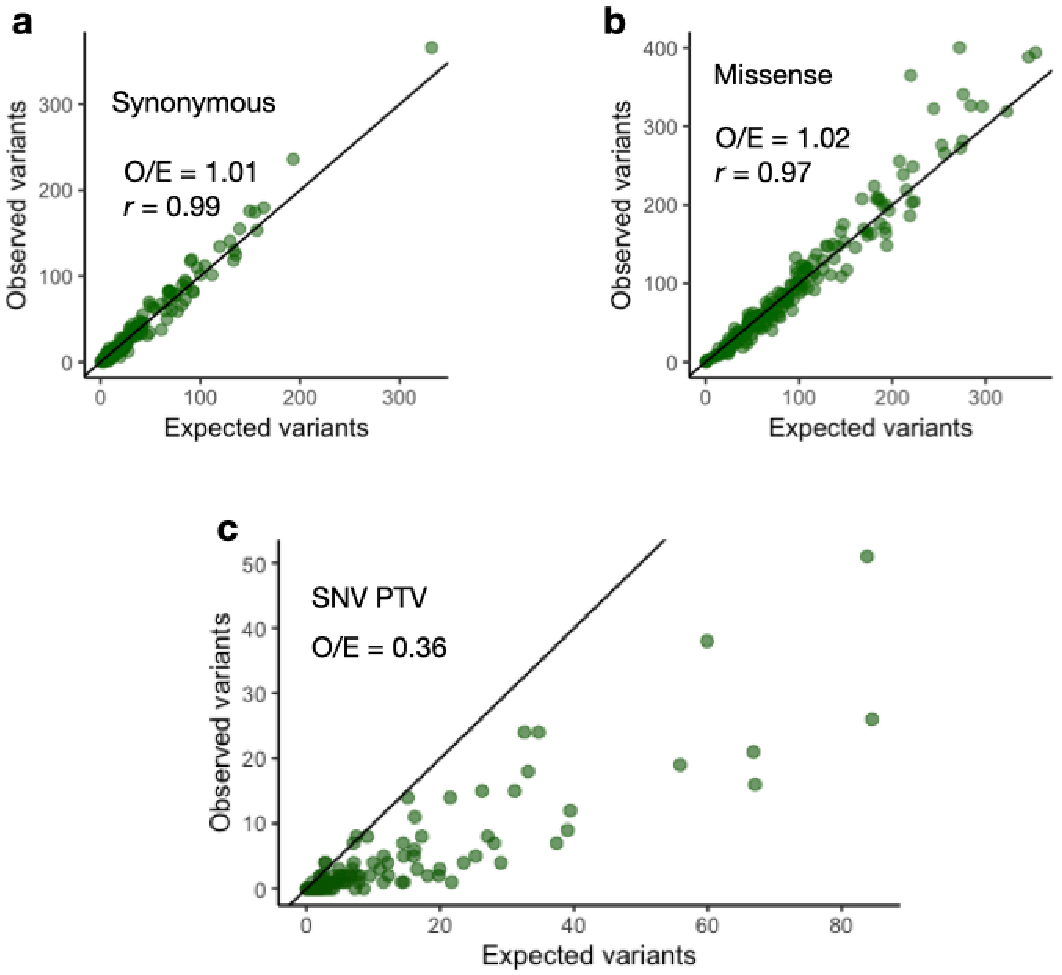


Figure 4.3 Validation of the constructed mutational model in *TTN* synonymous, *TTN* missense variants and TTN-truncating variants. For *TTN* synonymous and missense variants, the observed mutational burdens could be predicted from the mutational model. For TTN-truncating single nucleotide variants, the observed number is only about 1/3 of the expected number, verifying that they are under depletion. In the plot, each dot represents a type of substitution specified by trinucleotide sequence context and methylation level. There are 184 for non-CpG sites and 24 for CpG sites.

4.3.2 Identifying homologous residues intolerant of missense variants in I-set and FN3 domains of TTN

Narrowing down our search on I-set and FN3 domains, I evaluated the genetic constraint of missense variants in these two domains. Initially, I tested the idea of homologous residue constraint by only assessing domains in TTN before I extended the framework across all human proteins in Chapter 3. I would describe the original results of identifying TTN-specific constrained homologous residues here. Using the domain family alignments curated by the Pfam database, there are 164 and 132 copies for I-set and FN3 domains in TTN respectively. In terms of numbers of possible rare *TTN* missense variants (gnomAD AF<0.1% or unobserved), there are 91,036 in I-set domains and 72,012 in FN3 domains. Since rare missense variants could also act through cryptic splicing, I excluded 528 likely cryptic-splicing missense variants in TTN I-set and FN3 domains predicted by SpliceAI (delta score>0.5) in the calculation of missense constraint. As a whole, missense variants in I-set or FN3 domains of TTN are not under constraint (I-set: observed/expected=1.06, 95% CI=1.03-1.09; FN3: observed/expected=1.03, 95%CI=1.00-1.07).

To measure the genetic intolerance/constraint of *TTN* missense variants of residues in a homologous domain position, I calculated the depletion of variants. It is measured as the ratio of the total observed number of *TTN* missense variants over all residues (Observed) in the gnomAD reference population (v2.1.1 exome datasets 125,748 individuals) to the sum of TTN missense substitutions over all residues (Expected) expected in the gnomAD population predicted by the neutral mutational model described above. After filtering unqualified positions (number of expected variants <3, explained in **3.2.4**), both families have 94 homologous positions assessable out of 138 for I-set domains and 111 for FN3 domains.

The genetic constraint measured by the Observed/Expected ratio for each filtered homologous position was calculated and tested on significance. After correcting *P*-value in multiple testing to stringently limit the false discovery rate (Bonferroni adjusted *P*-value<0.05), there are three positions under significant constraint: I-set domain position 42, I-set domain position 117, and FN3 domain position 22 (**Table 4.1**).

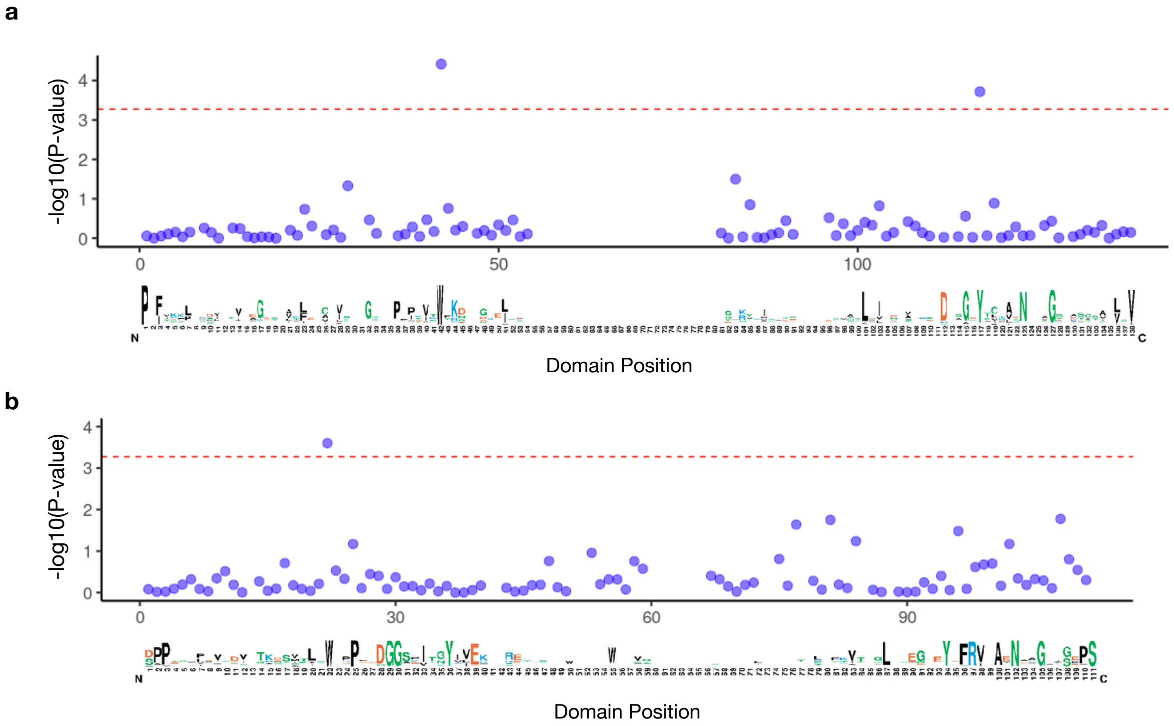


Figure 4.4 The *P*-value for homologous positions in TTN I-set and FN3 domains identified using a TTN-specific approach. The dot plots show the unadjusted P-value of constrained significance (Y-axis) along with the family homologous positions (X-axis). The red dashed line indicates the significance threshold for Bonferroni correction (P-value<0.05/94). Sequence logos are displayed at the bottom of P-value plots, which are generated from all domain's copies in TTN using https://weblogo.berkeley.edu. (a) The P-value of homologous residue constraint in I-set domain (b) The P-value of homologous residue constraint in FN3 domain.

Table 4.1 Significant constrained domain positions identified using TTN-specific approach

Domain Position	#Observed variants	#Expected variants	Observed/Expected ratio (95% CI)	Bonferroni adjusted *P*-value
I-set 42	44	76.5	0.57 (0.40-0.81)	3.6E-03
I-set 117	48	77.8	0.62 (0.44-0.86)	1.8E-02
FN3 22	33	58.1	0.57 (0.38-0.84)	2.4E-02

Missense variants at these three significantly constrained positions have been reported to be in association with *TTN*-related disorders. To be noticed, the only confirmed DCM-causing missense variant (p. Trp976Arg) is at the most constrained position (I-set domain position 42). On the second significantly constrained position (I-set domain position 117), it includes a *de novo* missense variant (p. Tyr3038His) reported at ClinVar associated with DCM.
The third most constrained position (fn3 domain position 22) also includes a known disease-associated residue, residue tryptophan at amino acid position 31729. Two missense variants at this position (p.Trp31729Cys and p.Trp31729Arg) have been reported previously in association with autosomal dominant hereditary myopathy with early respiratory failure (HMERF), which were suggested to impair the 119th-FN3 domain solubility and cause protein misfolding in functional studies[139]. Cardiac involvement was also observed in patients with HMERF carrying the mutation p.Trp31729Cys though without clear association with cardiomyopathies[140].

4.3.3 The significant constrained positions in I-set and FN3 domains in TTN are also constrained across human proteins

I compared the results generated by HRC considering all human proteins with the TTN-specific constraint analysis. There are 785 I-set domains in 186 human proteins (translated from RefSeq select transcripts) and 597 FN3 domains in 127 human proteins. 139 domain positions out of 220 are assessable in I-set domains and 115 out of 171 assessable in FN3 domains. For I-set domains across human proteins, there are 10 significant constrained positions (HRC<1, *P*-value<0.05) including the two positions (I-set 42 and I-set 117) identified in a TTN-specific approach described above (**Table 4.2**). Five significant constrained positions are identified in FN3 domains across human proteins including the one (FN3 22) identified in a TTN-specific approach (**Table 4.3**).

Compared with a TTN-specific approach, the proteome-wide approach implemented in HRC not only confirmed the same significant positions with narrower confidence intervals (**Figure 4.5**) but also was able to prioritise 12 more significant positions with greater statistical power. In total, there are 14,964 all possible rare missense variants in these positions. To be noticed, the recently reported novel DCM-associated missense variant Cys3892Ser[138] is found at the I-set 119 position, which is one of the significant constrained positions identified by the proteome-wide approach. So far, the two DCM-causing *TTN* missense variants (reported with strong linkage evidence) could be both mapped to our prioritised domain positions.

Table 4.2 Significant constrained homologous positions of I-set domains identified across human proteins.

Domain Position	#Observed missense variants	#Expected missense variants	Observed/Expected ratio (95% CI)
I-set 42	183	340.1	0.53 (0.46-0.60)
I-set 117	242	338.5	0.71 (0.62-0.80)
I-set 119	307	395.0	0.78 (0.68-0.85)
I-set 26	330	414.0	0.80 (0.70-0.88)
I-set 127	330	407.4	0.81 (0.71-0.90)
I-set 115	415	507.6	0.82 (0.72-0.89)
I-set 101	227	268.8	0.85 (0.72-0.98)
I-set 23	254	296.4	0.85 (0.74-0.99)
I-set 103	333	388.6	0.86 (0.75-0.97)
I-set 32	387	450.2	0.85 (0.77-0.97)

Table 4.3 Significant constrained homologous positions of FN3 domains identified across human proteins.

Domain Position	#Observed missense variants	#Expected missense variants	Observed/Expected ratio (95% CI)
FN3 22	117	247.0	0.47 (0.39-0.54)
FN3 108	173	238.2	0.73 (0.61-0.83)
FN3 94	204	264.9	0.77 (0.65-0.87)
FN3 87	164	203.9	0.80 (0.67-0.96)
FN3 36	202	247.0	0.82 (0.69-0.96)

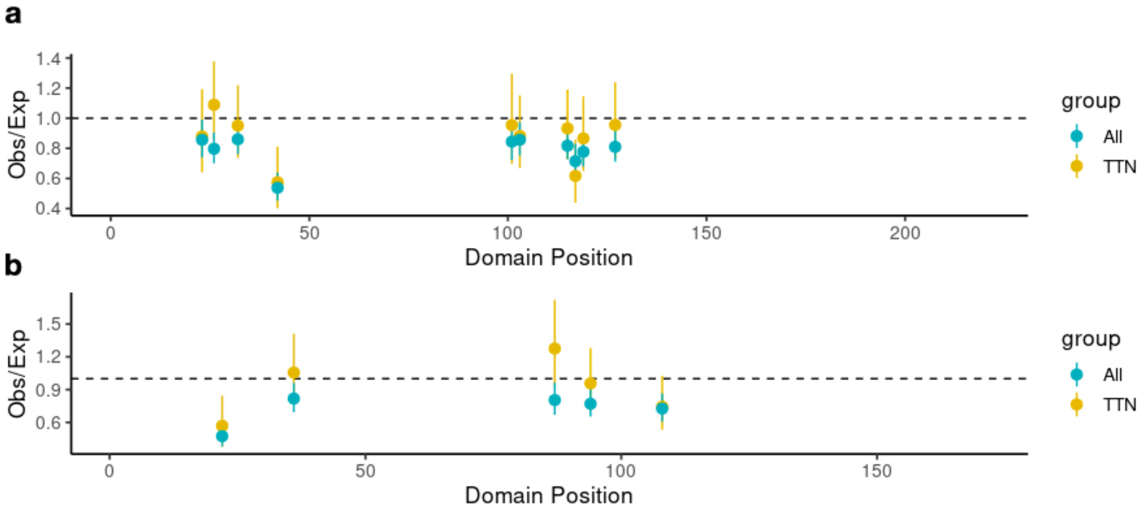


Figure 4.5 Compare the significant constrained homologous positions of I-set and FN3 in TTN and across domains. The genetic constraint measured by Obs/Exp ratio and its 95% CIs are shown for **(a)** I-set and **(b)** FN3 domains for both exclusively within TTN and across human proteins.

4.3.4 Significant constrained positions are enriched with known disease-associated variants both in *TTN* and other genes

I ask whether the missense variants affecting constrained positions are enriched with known disease-associated variants. From ClinVar, there were 209 ClinVar Pathogenic/Likely pathogenic and 644 Benign/Likely benign missense variants in I-set and FN3 domains of 85 human proteins. Among them, *TTN* has 20 Pathogenic/Likely pathogenic variants and 375 Benign/Likely benign variants.

Now we have two lists of constrained positions: one identified by the TTN-specific approach and the other by the proteome-wide approach. The TTN-specific list is a subset of the proteome-wide list. I am going to conduct three tests to answer the question: (1) whether constrained domain positions by TTN-specific approach are associated with the 20 *TTN* pathogenic ClinVar variants; (2) whether constrained domain positions by proteome-wide approach are associated with the 20 *TTN* pathogenic ClinVar variants; (3) whether constrained domain positions by proteome-wide approach are associated with the 209 pathogenic ClinVar variants in all 85 human genes. For each test, the proportion of pathogenic variants (Prop$_{pathogenic}$) is compared with the proportion of benign variants (Prop$_{benign}$) in constrained positions. A one-sided binominal test is applied to compare the difference of two proportions. Alternatively, the odds ratio is also reported, which measures whether variants affecting constrained positions are more likely to be pathogenic versus benign.

Compared with benign variants, pathogenic *TTN* ClinVar variants are enriched in the constrained domain positions identified using either the TTN-specific approach (**Table 4.4**: Test1, *P-value*<0.05) or the proteome-wide approach (**Table 4.4**: Test2, *P-value*<0.05). The association was confirmed in both I-set and FN3 domain. The constrained positions are also significantly associated with (likely) pathogenic ClinVar variants across all 85 human genes in either I-set domains or FN3 domains (**Table 4.4**: Test3, *P-value*<0.05).

Table 4.4 Significant tests to assess whether constrained domain positions are associated with pathogenic ClinVar variants.

	Test 1: whether constrained domain positions (TTN-specific approach) are associated with the 20 *TTN* pathogenic ClinVar variants			
	% pathogenic variants at constrained positions ($\text{Prop}_{\text{pathogenic}}$)	% benign variants at constrained positions ($\text{Prop}_{\text{benign}}$)	Binomial test *P*-value of $\text{Prop}_{\text{pathogenic}}$ vs $\text{Prop}_{\text{benign}}$	*Odds Ratio (ratio of odds of pathogenic variants in constrained versus unconstrained positions)* and 95%CI
I-set	2/11	4/213	0.02	11.6 [1.9-71.9]
FN3	2/9	1/162	0.001	46.0 [3.7-570.0]
I-set and FN3	4/20	5/375	1E-04	18.5 [4.5-75.5]
	Test 2: whether constrained domain positions (proteome-wide approach) are associated with the 20 *TTN* pathogenic ClinVar variants			
I-set	4/11	19/213	0.01	5.8 [1.6-21.7]
FN3	2/9	6/162	0.04	7.4 [1.3-43.6]
I-set and FN3	6/20	25/375	2E-03	6.0 [2.1-17.0]
	Test 3: whether constrained domain positions (proteome-wide approach) are associated with the 209 pathogenic ClinVar variants in all 85 human genes			
I-set	37/125	29/376	5E-13	5.0 [2.9-8.6]
FN3	13/84	6/268	5E-08	8.0 [2.9-21.8]
I-set and FN3	50/209	35/644	4E-12	5.5 [3.4-8.7]

4.3.5 Case-control burden test suggests constrained *TTN* missense variants in I-set domains are associated with subsets of DCM cases

To evaluate the association between constrained *TTN* missense variants and DCM, I attempted to conduct a case-control analysis to compare the enrichment of constrained *TTN* missense in patients of DCM versus healthy controls. I collected 2,023 patients of DCM and 2,313 healthy volunteer controls without cardiac conditions confirmed by cardiac MRI. As individuals in these cohorts were collected from three different sites (UK Royal Brompton Hospital, Egypt Aswan Heart Centre, and Singapore National Heart Centre) with different ethnicities, joint variant calling was performed with the same quality control criteria. As samples of different ethnicities/sites show a different prevalence of missense variants (**Table 4.5**), I conduct the burden tests in separate ethnicities and use the site- and ethnicity-matched controls including Caucasian samples (CAU), African samples (AFR), and East Asian samples (EAS). The ORs for DCM for carriers of *TTN* missense variants, synonymous variants, and TTNtv are shown as reference in **Figure 4.6**. South Asian samples are excluded in the following analysis due to low sample size. 13 missense variants in the samples predicted as cryptic splicing (SpliceAI delta score>0.5) are excluded in the downstream analysis.

Constrained variants identified by proteome-wide methods were tested on their association with DCM as there are more constrained variants identified providing higher statistical power. While there is no enrichment detected by testing all constrained *TTN* missense variants together for both I-set and FN3 domains in any of the ethnicity groups (**Figure 4.7a**), the constrained variants in I-set domains shows significant association with DCM cases (OR=3.18, 95% CI= 1.30-9.34, *P*-value=0.006) in the Caucasian case-control burden test. For the burden tests in African and East Asian cohorts, there is no significant association between constrained variants in I-set domains and DCM cases (OR_{AFR}=0.15, 95% CI=0.004-1.00; OR_{EAS}=1.18, 95% CI = 0.22-4.16). The burden tests in constrained variants in FN3 domains did not show any significant signal in any of the ethnicity groups (**Figure 4.7a**).

Next, I tested the enrichments of constrained variants in I-set domains in separate DCM cohorts in Caucasian samples using the same Caucasian control cohort. There are three subgroups in the DCM cohorts: familial DCM (n=898), alcoholic cardiomyopathy (ACM; n=141) and chemotherapy-induced DCM (Chemo; n=108). Significant association is identified in ACM-control burden test (OR=5.69, 95% CI=1.77-18.77). For familial DCM and Chemo cohorts, a trend of increased association was observed compared with unconstrained variants. For FN3 domains, no association was found in any of the DCM subgroups.

Table 4.5 Prevalence (rate per individual) of rare (MAF<0.1% or unobserved in gnomAD) missense and synonymous variants in cohorts of patients with DCM and healthy controls

Ethnicity (Self-reported)	Site	Cohort	Cohort size	Prevalence of Missense Variants	Prevalence of synonymous variants
Caucasian	UK Royal Brompton Hospital	Case	1,147	0.40	0.20
Caucasian	UK Royal Brompton Hospital	Control	671	0.42	0.21
African	Egypt Aswan Heart Centre	Case	124	0.65	0.39
African	Egypt Aswan Heart Centre	Control	515	0.61	0.40
East Asian	Singapore National Heart Centre	Case	101	0.68	0.31
East Asian	Singapore National Heart Centre	Control	713	0.64	0.37
South Asian	Singapore National Heart Centre	Case	96	0.63	0.41
South Asian	Singapore National Heart Centre	Control	25	0.60	0.36

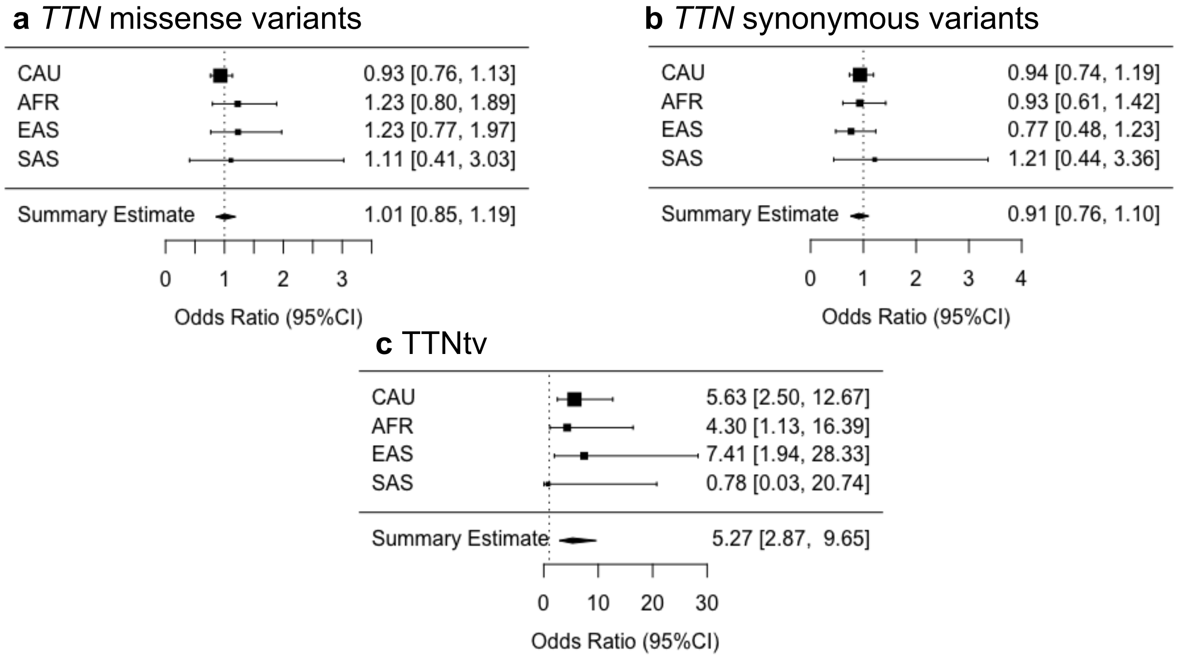


Figure 4.6 Burden test of *TTN* variants in patients with DCM versus healthy controls. *TTN* missense **(a)**, *TTN* synonymous variants **(b)**, and TTN-truncating variants **(c)** are tested. *TTN* missense and *TTN* synonymous variants collectively don't show significant association with DCM while TTNtv is significantly associated with DCM. Four site- and ethnicity-matched burden tests were included: Caucasian samples (CAU) from UK Royal Brompton Hospital (N_{DCM}=1,147, $N_{control}$=671), African samples (AFR) from Egypt Aswan Heart Centre (N_{DCM}=124, $N_{control}$=515), East Asian samples (EAS) from Singapore National Heart Centre (N_{DCM}=101, $N_{control}$=713) and South Asian samples (SAS from Singapore National Heart Centre (N_{DCM}=96, $N_{control}$=25). The pooled effect size was derived using meta-analysis under a fixed-effect model.

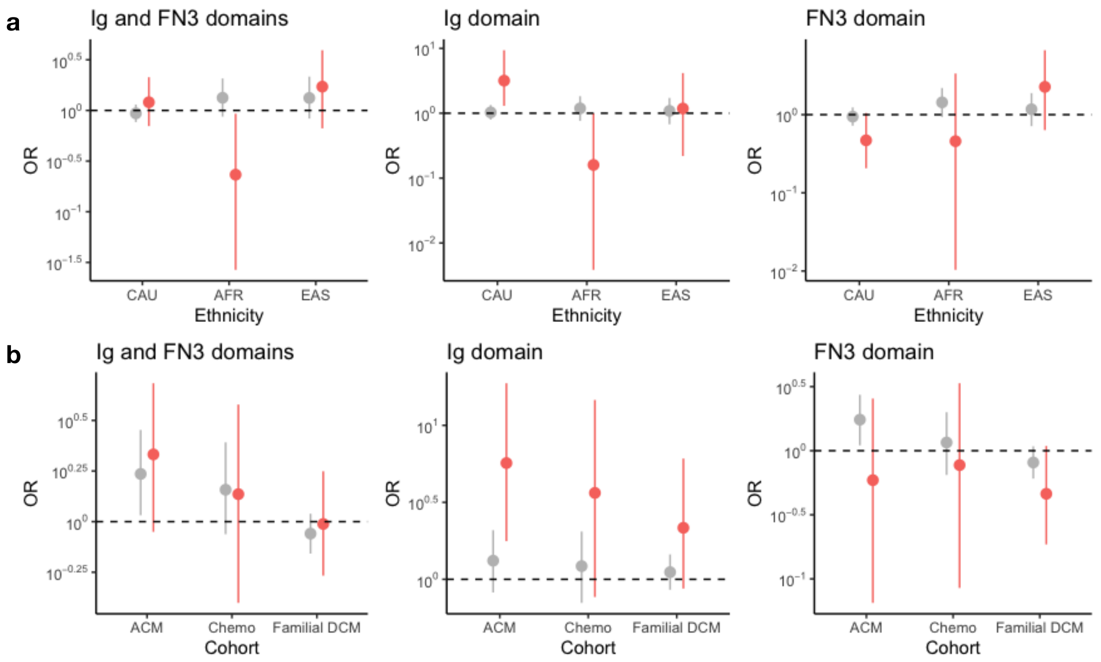


Figure 4.7 Burden tests of *TTN* unconstrained (grey) and constrained (red) missense variants in DCM cases versus controls. Variants in Ig domains and FN3 domains are tested both as a whole and separately. **a.** Site- and ethnicity-matched burden tests on DCM cases versus controls. The error bars represent 95% CIs of odds ratio on $\log_{10}$ scale. **b.** Burden tests on subsets of DCM cases versus controls in Caucasian samples.

4.4 Discussion

4.4.1 Strengths and limitations of the study

In this Chapter, I have evaluated an approach to identify constrained *TTN* missense variants in I-set and FN3 domains. Across human proteins, 15 domain positions are significantly intolerant of missense variants in I-set and FN3 domains. Three of them also reached a significance level when only examining homologous residues in TTN. In *TTN* I-set and FN3 domains, these 15 positions have 14,964 possible rare missense variants. In the validations, the constrained positions are significantly associated with ClinVar pathogenic variants in either I-set or FN3 domains. The two reported DCM-causing missense variants with strong linkage evidence so far occurred in these top constrained positions. In site- and ethnicity-matched case-control burden tests, missense variants disrupting these constrained positions are significantly enriched in cases of DCM compared with healthy controls in Caucasian samples (OR=3.18, 95% CI= 1.30-9.34). When I analysed the same Caucasian DCM cohorts separately, this significant association is also found in the ACM cohort (OR=5.69, 95% CI=1.77-18.77). This finding presents first-time evidence based on population data to suggest the link between *TTN* missense variants and DCM. It also highlights using homologous residue constraint to prioritise underrecognized missense variants as an effective approach.

While we assessed homologous residues in I-set and FN3 domains, there are subsets of missense variants not considered in the study including the ones from positions underpowered in I-set and FN3 domains and other regions of *TTN*. Meanwhile, the effect of specific residues or positions in TTN might deviate from the average of all human proteins. As our approach measured intolerance of heterozygous missense variants, it might not be sensitive to prioritise mutation targets in the homozygous or compound heterozygous state.

As DCM usually occurs in adults, mutation targets might not be well-captured by purifying selection signals.

4.4.2 Discussion on future work

Though the findings so far are promising, there are more questions unanswered to understand the link between *TTN* missense variants and DCM.

How to replicate the findings?

To avoid false positives, we need to conduct independent validations. It can also help to understand the heterogeneous association levels observed in the study. Apart from comparing cases versus healthy controls, we could also compare cases with DCM versus cases without heart muscle diseases. To increase statistical power and understand the phenotypic effects, the association between sub-clinical phenotypes and outcomes could also be evaluated. Further evaluation could make use of UK Biobank with the collection of exome sequencing, cardiac imaging, and clinical outcome data. Among 200,628 participants with whole-exome sequencing data of recent UKBB release (Oct 2020), there are 21,322 of them with Cardiac MRI data, 341 of them with DCM, and 3475 of them with the outcome of heart failure.

We should also validate the findings in populations of different ancestries. In our current case-control burden test, no association was detected in non-Caucasian populations, which might be limited by sample size or different genetic architectures. Up to date, there haven't been studies on DCM in non-Caucasian populations with a sample size comparable to studies in Caucasians. Another reason is due to the fact that our signal mainly comes from mutation pattern in Caucasian populations, which has lower effective population size compared with African populations. Thus, the constrained signals we discovered might have low sensitivity when applying in other populations. Taking these factors together, their interaction likely explains the results we saw in non-Caucasian populations.

What's the molecular mechanism of DCM-causing *TTN* missense variants?

From the two reported DCM-causing *TTN* missense variants with linkage evidence, they both likely act through a loss-of-function mechanism by destabilising immunoglobin domains. To find clues about the new *TTN* missense variants, before carrying functional studies we could also gain insight to predict the function of residues and the structural effects of missense variants through structural modelling. Existing resources could support the structural modelling such as domain structural models curated by Pfam[124], structural models for TTN domains (TITINdb)[141], and tools predicting structural effects of missense variants (e.g. missense3D[33]). Compared with gain-of-function variants, loss-of-function variants are more likely to be in the hydrophobic core of the protein to disrupt protein stability[142]. A preliminary analysis indicates that the top constrained positions in I-set and FN3 domain have a conserved tryptophan residue, which is buried in the hydrophobic core. Further analysis on other residues is needed to understand the underlying disease mechanism.

Is *TTN* missense in FN3 domains not associated with DCM?

In our study of case-control burden tests, we did not find any significant association between constrained missense variants in FN3 domains and DCM. For all ClinVar (likely) pathogenic variants in *TTN* FN3 domains, they are all reported to be associated with myopathy disorders. As already mentioned above, a study assessing the cardiac phenotypes of 22 HMERF patients with p. Trp31729Cys, a missense variant disrupting the most constrained position in FN3 domain (FN3 22), did not find a clear association between the variant and cardiac conditions. While the above observation cannot exclude an association, we hope to have higher confidence in answering this question by assessing the carriers' sub-clinical cardiac phenotypes and outcomes in cohorts with a larger sample size.

4.4.3 Conclusion

Overall, our study prioritises a list of promising candidate disease-associated *TTN* missense variants for follow-up studies. It also emphasizes that *TTN* missense variants should not be ignored in clinical genetic diagnosis.

4.5 Outline of further work

- Replicate the burden tests using a list of constrained variants identified with more stringent level significance (e.g*., P-value* < 0.01).
- Study the association of constrained variants with cardiac phenotype and clinical outcomes using UKBB data.
- Predict the structural effect of constrained variants on domains
- Explore whether integrating genetic constraint with exon expression in cardiac tissues, conservation, and structural effect prediction could further improve the separation of DCM-associated variants from benign ones.

4.6 Acknowledgements

- Pantazis Theotokis: ran the bioinformatic pipelines on variant calling and quality control jointly for multiple samples from different sites.
- Rachel Buchan: performed sample sequencing on cases of DCM and controls
- Deborah Morris-Rosendahl and Shibu John: curated the samples of DCM from the RBH clinical diagnostic laboratory.
- Stuart Cook and Chee Jian Pua: curated the samples of DCM cases and healthy controls from the Singapore National Heart Centre.

- Magdi H. Yacoub, Yasmine Aguib, and Mona Alouba: curated the samples of DCM cases and healthy controls from Egypt Aswan Heart Centre
- Nicola Whiffin: study design; overall mentorship
- James Ware: study design; overall mentorship

Chapter 5 Annotating high-impact 5’UTR variants with UTRannotator

5.1 Introduction

On a strand of mRNA, a proportion, either from its 5’ end to the start codon of a coding sequence or from the 3’ end to the stop codon is untranslated, known as untranslated regions (UTRs). Untranslated regions play an important role in regulating gene expression at the post-transcriptional level. This regulation by UTR is mediated through UTR elements of mRNA including upstream open reading frames (uORF), sequence motifs and secondary structures[143]. Upstream ORF is defined by a start codon in the 5’ untranslated region. About 49% of human transcripts with 5’UTRs have at least one uORF[144]. In translation initiation, since the ribosomes would begin scanning the messenger RNA from 5’unstranlated regions, the presence of an upstream start codon preceding the main coding sequence (CDS; protein-coding ORF) could also initiate translation.

Not every ORF is translated. The propensity of a start codon initiating translation depends on the similarity of its local sequence context to the “Kozak consensus”. “Kozak consensus” sequence describes the most conserved sequence pattern of protein translation initiation sites in eukaryotic mRNA transcripts[145]. When a uORF is flanked with unfavourable sequence (weak Kozak context), leaky scanning could occur such that scanning ribosomes would skip the uORF[143].

Upon translation initiation, upstream ORFs can affect phenotypes via different mechanisms: (1) uORF translation could inhibit the translation efficiency of protein-coding ORF. For some genes, after terminating uORF translation, ribosome subunits dissociate from mRNA thus reduce or delay the protein expression of CDS. For other genes, the interruption could also be mediated through uORF-encoded peptides, which can directly interact with and stall the scanning ribosomes. As a consequence, uORF translation imposes a physical barrier to ribosomes getting access to main ORFs. Overall, it has been shown that naturally occurring uORF can reduce downstream protein translation by 30-80%[144]; (2) the translated micropeptides from uORFs could also have distinct biological functions from the CDS-encoded protein. In a recent systematic study[146], uORF-encoded microproteins are suggested to have critical roles in cellular growth. They could also have distinct cellular localization and even form stable complexes with the CDS-encoded protein at the same messenger RNA.

Depending on the presence of stop codons, a uORF can be categorised into different subtypes: (a) a stringent upstream ORF which also has a stop codon in the 5'UTR (**Figure 5.1**a);   an overlapping ORF (oORF) which has the stop codon in the main coding sequence thus the uORF is overlapping with the CDS; (b) If the oORF is not in the same frame of the main coding sequence (i.e. the distance between uAUG and downstream AUG of CDS is a multiple of three nucleotides), it's called out-of-frame oORF (**Figure 5.1**b); (c) otherwise, it's in-frame oORF (**Figure 5.1**c);

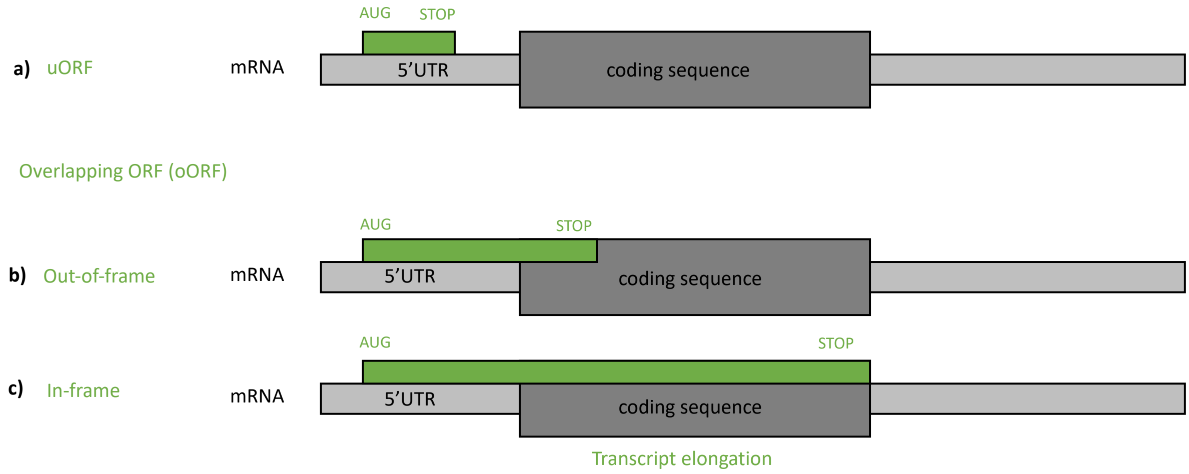


Figure 5.1 The subtypes of upstream ORFs. (a) stringently defined uORF with the entire ORF located in 5'UTR. (b) Out-of-frame Overlapping ORF (c) In-frame Overlapping ORF. Reproduced from Nicola Whiffin with permission.

Genetic variants creating or disrupting uORFs (uORF-perturbing variants) were under-recognised in previous human genetic studies. Up till May 2020, there were only 97 5'UTR variants interpreted as Pathogenic/Likely pathogenic in ClinVar. While these isolated cases were reported in previous studies, Whiffin *et.al*[31] conducted a systematic study to show that uORF-perturbing variants can cause human diseases by leveraging 15,708 gnomAD whole-genome sequences. It has been shown that genetic variants introducing or disrupting uORF are under strong negative selection. Several categories of factors found in uORF-perturbing variants are under stronger selection thus more likely to cause diseases including variants with strong Kozak context of upstream start codons, perturbing a uORF overlapping with the main coding sequence, disrupting a uORF with translation evidence, or occurring in loss-of-function intolerant genes.

However, there wasn't a bioinformatic tool available to annotate this specific class of variants. To help rapidly identify this important class of disease-causing variants, I have developed a plugin called UTRannotator to use with the mainstream variant annotation tool, Ensembl VEP[4].

UTRannotator can identify whether any small variation (1-5bp) is uORF-perturbing including: variants creating a new upstream start codon, removing start codons or stop codons of an existing uORF, creating an earlier stop codon for an existing uORF, or shifting the frame of an existing uORF. The tool would also output detailed annotations relevant to evaluate the impact of the uORF. When a variant disrupts multiple uORFs, all possible consequences and annotations would be reported.

To evaluate the clinical utility of UTRannotator, I applied the tool in the ClinVar dataset[72], gnomAD[20], and Genomics England 100K genomes project[147]. I demonstrate that UTRannotator can detect high-impact disease-associated variants for further experimental validation.

5.2 Method

5.2.1 Implementation and Code availability

The plugin is developed in Perl. By making use of Ensembl Perl modules, it can quickly get access to necessary info for 5'UTR variant annotation including genomic descriptions of variants, gene and transcript annotation with exon/intron boundaries, and UTR sequence from Ensembl database. It is available on Github (https://github.com/ImperialCardioGenetics/UTRannotator).

Instructions about install and run UTRannotator is in the Github repository (*https://github.com/ImperialCardioGenetics/UTRannotator/blob/master/Supplementary_Information.pdf*). It could take an input list of variants with the following formats including default

VEP input format, VCF, HGVS, and variant identifiers recognised by Ensembl database. For the output format, it supports default VEP output format, tab-delimited output and VCF output.

5.2.2 Consequence annotated

For any SNV, 1-5bp small insertion/deletion (indel) or multi-nucleotide variant in a 5’ UTR, the UTRannotator would first summarize the number of uORFs in the 5’UTR in the reference sequence. Then, for each variant within the 5’UTR, the tool evaluates whether it would have any of the following consequences, on any annotated transcript: (1) creating a new start codon AUG to introduce a new uORF; (2) removing an existing start codon AUG; (3) removing the STOP codon of an existing uORF; (4) creating a new stop codon to shorten an existing uORF; (5) disrupting an existing uORF with a frameshift deletion or insertion, whose number of nucleotides inserted or deleted is not a multiple of three. Where a variant has multiple annotation consequences, it is evaluated for each separately.

5.2.3 Detailed annotation for each consequence

To enable evaluation of the effect of each variant, the UTRannotator outputs detailed annotations for each type of uORF-perturbing variant (**Table 5.1**). This includes describing the subtype of uORF created and/or disrupted (i.e. whether this is a distinct uORF with a stop codon in the 5’UTR, or an ORF that overlaps the coding sequence either in- or out-of-frame), and the strength of the created and/or disrupted uORF start site match to the Kozak consensus sequence[145]. For a variant disrupting an uORF, the tool also evaluates whether the uORF has any experimental evidence of translation, by assessing translated uORFs previously identified in ribosome profiling experiments. We downloaded a list of translated uORFs curated by the public online repository for sORFs (sorfs.org)[148] aggregating from existing experimental data. Users can also use their own customised list of translated uORFs. Given that ribosome profiling datasets are currently limited in the cell types/tissues and conditions analysed, the tool outputs results for all possible uORF-disrupting variants and includes experimental evidence as an annotation.

Since a 5'UTR can have multiple existing uORFs, for each 5'UTR variant the UTRannotator outputs the annotations for all disrupted uORFs.

Table 5.1 Details of the annotations provided for different categories of uORF-perturbing variants

Consequence	uAUG-gained	uAUG-lost	uSTOP-lost	uSTOP-gained	uFrameshift
Number of existing uORFs	√	√	√	√	√
KozakContext: sequence and strength	√	√	√	√	√
Start distance to CDS	√	√		√	√
Start distance to STOP	√	√			
With translated evidence		√	√	√	√
uORF subtype	√	√		√	√ (ref and alt)
Other Annotations	Start distance from cap		Whether there is an alternative STOP, alternative stop distance to CDS, frame of disrupted uORF with CDS	New stop distance to CDS	

5.2.4 Time complexity to run UTRannotator

The time complexity of this implementation is linear to the number of input variants. The ratio of running time without the plugin to that with the plugin, tested on 1000 random variants (60% annotated as 5'UTR variants) is 1.02-1.07 (5 replications).

5.3 Results

5.3.1 Application on ClinVar

To show the utility of our UTRannotator tool, I annotated all 5'UTR variants interpreted as pathogenic/likely pathogenic and uncertain significance from ClinVar (version 202005) (Landrum et al., 2018). These variants do not have a coding annotation on any transcript. However, I note that 5'UTR variants are under-represented in ClinVar as they are rarely sequenced and/or reported.

There are 97 Pathogenic/Likely pathogenic 5'UTR variants in ClinVar (97/113,969=0.085% of all ClinVar Pathogenic/Likely pathogenic). 91 are 1-5bp small variations, 29 of which (31.9%) are annotated as creating or disrupting uORFs by our plugin (**Figure 5.2**). I examined the evidence behind the reported clinical significance for each variant, and found 15 (51.7%) have previously been attributed to a uORF-perturbing mechanism.

There are 5,128 5'UTR variants of uncertain significance (VUS) reported in ClinVar (5,128/255,691=2% of all VUS), 4,966 of which are 1-5 bp small variations. The plugin annotated 377 of these (7.6%) as creating or disrupting uORFs, on at least one annotated transcript.

The detailed annotations from the UTRannotator can be used to illustrate how to prioritise high-impact 5’ UTR VUS that are most promising for further follow-up. Variants were first restricted to the ones that form new overlapping ORFs (oORFs) with start sites that are Strong or Moderate matches to the Kozak consensus sequence, or that are uORFs with documented evidence of translation, as it has been previously shown that variants with these consequences are under strongest negative selection[31]. Finally, we selected variants in 3,191 genes previously categorised with a ‘High’ likelihood to operate through uORF-perturbation mechanism[31]. Through this approach, 31 potential ‘high-impact’ ClinVar 5’UTR VUS could be identified.

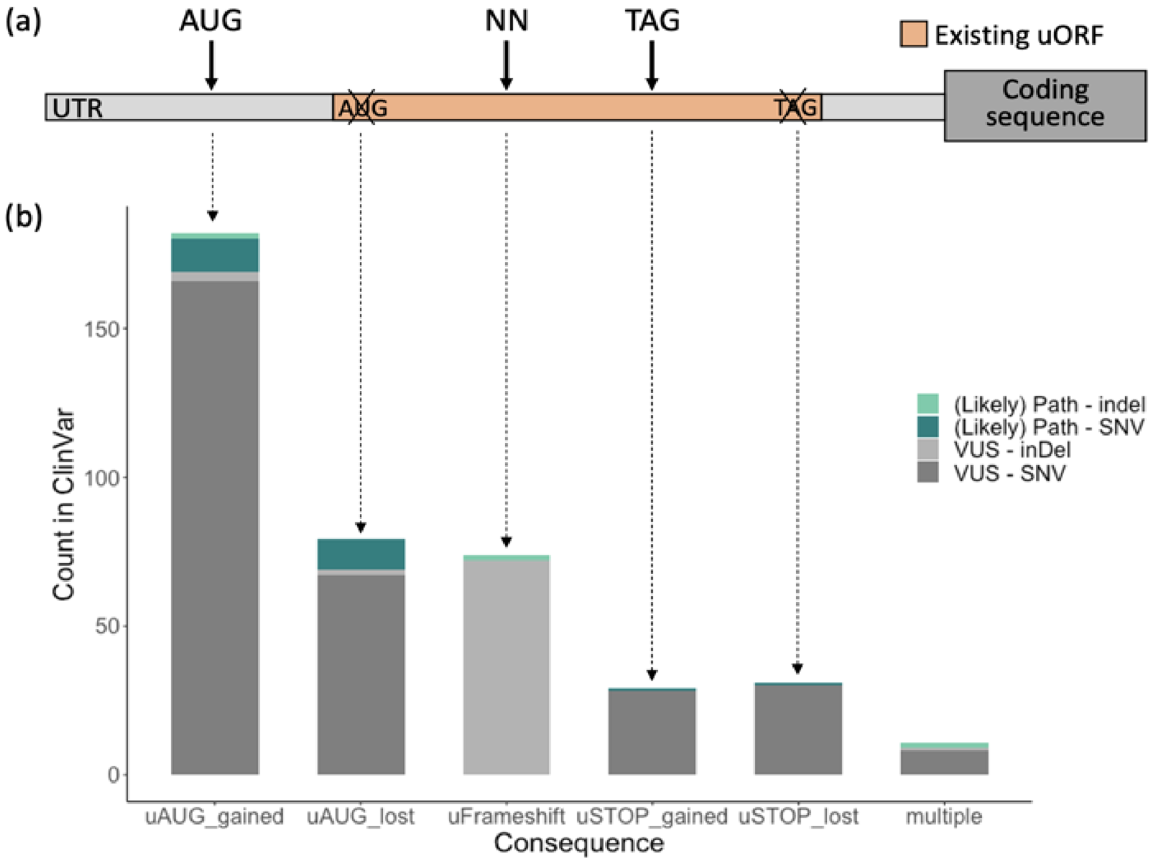


Figure 5.2 5’UTR variants in ClinVar annotated by the UTRannotator. (a) A schematic showing the five distinct consequences of 5’UTR variants annotated by the tool: those that create an upstream AUG (uAUG_gained), those that disrupt the start site of an existing upstream open reading frame (uORF; uAUG_lost), those that cause a frameshift in the sequence of the uORF (uFrameShift), those that introduce a new stop codon into an existing uORF (uSTOP_gained) and those that disrupt the stop site of an existing uORF (uSTOP_lost). (b) The counts of each variant category that are classified as Pathogenic/Likely Pathogenic (teal) or Uncertain Significance (VUS; grey) in ClinVar.

5.3.2 Application on gnomAD and Genomics England

To study the potential contribution of uORF-perturbing variants to undiagnosed cases, I compared its burden in undiagnosed cohorts from Genomics England 100K Genomes Project (GEL)[147] with that in general populations from gnomAD. I included *de novo* variants (version 2020 Sept) of rare disease probands from GEL. GEL applies a Rare Disease Tiering process to annotate plausibly pathogenic variants. In order to remove probands with any potential protein-coding diagnostic variants, I excluded any trios in the downstream analysis if the proband has either Tier 1 variants (protein-truncating variants) or Tier 2 variants (protein-altering variants) on known disease gene panels applied in the participants. In total, there were 859,350 *de novo* variants from 12,456 trios annotated in the downstream analysis. As a control dataset, 5'UTR variants from gnomAD v3 71,702 genomes were also included in the analysis.

UTRannotator was applied in these two datasets to analyse the burden of uORF-perturbing variants in predicted loss-of-function (LoF) intolerant genes (defined by gnomAD gene constraint metric LOEUF<0.35[20]). LOEUF is a conserved estimate indicating how much a gene is under depletion of heterozygous loss-of-function variants. A lower value suggests a higher intolerance. 0.35 is a recommended threshold by authors. There are comparable proportions of 5'UTR variants in the two datasets (proportions of people with at least one variant: GEL: 0.36%; gnomAD: 0.39%). As shown in **Figure 5.3**, 9.7% (66/683) of the 5'UTR variants are uORF-perturbing in the GEL *de novo* variant dataset, while there are only 7.1% (35,075/494,364) 5'UTR variants annotated as uORF-perturbing in gnomAD dataset (one-tailed binomial test *P-value*=0.005). For high-impact uORF-perturbing variants (variants creating/disrupting an oORF with strong Kozak context or disrupting a uORF with translation evidence), there are 1.5% (10/683) GEL *de novo* 5'UTR variants annotated as high-impact uORF-perturbing, as it's compared to only 0.2% (1,111/494,363) of gnomAD 5'UTR variants in the same category (one-tailed binomial test *P-value*=1.7×10^{-6}). This analysis shows that *de novo* 5’UTR variants from undiagnosed probands are enriched with high-impact uORF-perturbing variants in LoF-intolerant genes compared with general populations.

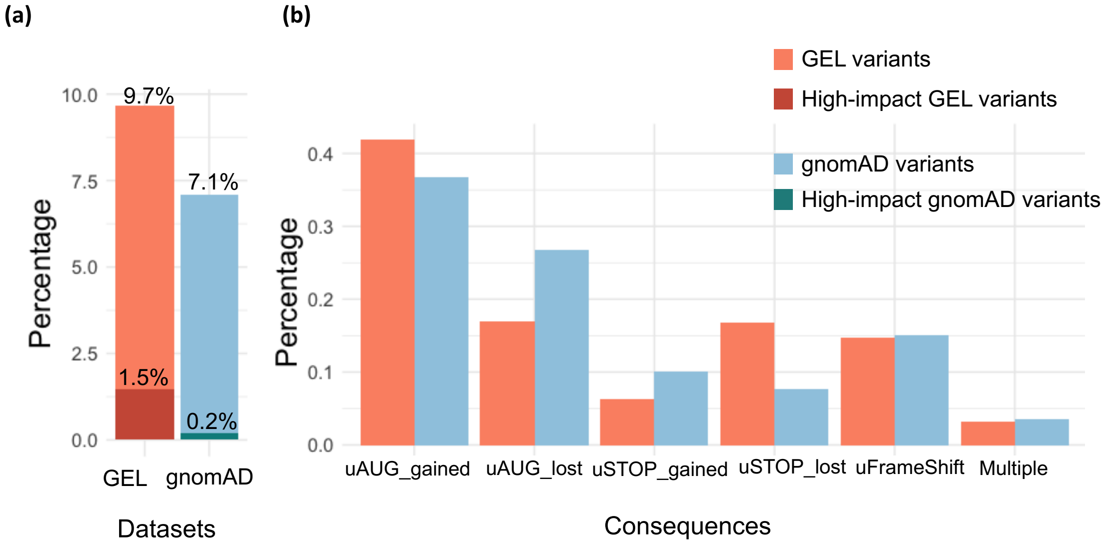


Figure 5.3 uORF-perturbing variants in loss of function intolerant genes found in GEL de novo variants and gnomAD variants. (a) the percentage of (high-impact) uORF-perturbing variants in 5’UTR; (b) the distribution of consequences in uORF-perturbing variants.

5.4 Discussion

5.4.1 Strengths of the study

I have developed a freely available tool, as a plugin to the Ensembl VEP, which annotates variants that create or disrupt uORFs. The output from the tool can be used to evaluate the possible impact of 5'UTR variants identified in patients for a role in disease. It is also directly applicable to annotate 5'UTR variants from other eukaryotes. To test its utility, we found that 31 ClinVar 5'UTR VUS are uORF-perturbing of high-impact, which could be reinterpreted in the future. We also found a significant enrichment of high-impact uORF-perturbing variants in undiagnosed cohorts from GEL 100K compared with the gnomAD population.

In both applications, as we see, the majority of the uORF-perturbing variants are the ones creating new upstream ORFs. This could be explained by the larger number of possible nucleotide positions to create new uAUGs compared to other uORF-perturbing variants. What's more, given that naturally occurring uORFs are selected to be short and far upstream of CDS, it hints to us that a subset of *de novo* variants creating a new uORF could be under strong selection. This could help to explain the higher burden of uAUG-creating variants we saw in GEL 100K undiagnosed patients compared with gnomAD (**Figure 5.3**b).

5.4.2 Limitations of the study and discussion for future directions

There are several limitations to the UTRannotator. I will also discuss and suggest future research directions.

Firstly, the UTRannotator only considers variants up to 5bps in length. The length is restricted due to two reasons: (1) the annotation of longer indels is difficult, as the chance of variants having multiple possible annotations is increased, and (2) the impact of larger indels that add or remove large stretches of UTR is currently unclear.

It also currently only considers uORFs with canonical AUG start sites. It is known that many translated uORFs use non-canonical start sites[149]. More research is needed into the impact of variants that create or disrupt these non-canonical uORFs in human disease.

While we focus on small variants of loss-of-function effects, the remaining UTR variants are largely uninterpreted. For SNVs within naturally occurring uORFs that do not change the length of uORFs, they are not under selective constraint collectively[150]. However, additional types of deleterious UTR variants could be added to the UTRannotator when they are discovered in new studies.

In the study, I show that by using UTRannotator, undiagnosed probands of rare diseases carry more high-impact *de novo* uORF-perturbing variants in LoF-intolerant genes compared to that with control populations. However, here it's still unclear whether uORF-perturbing variants are solely associated with *de novo* variants or associated with disease status. To disentangle the two explanations, as future work we could collect *de novo* non-coding variants from control trios (e.g. whole-genome sequencing data of unaffected siblings from the Simons Simplex Collection[151]) and compare them with that from trios of rare diseases. We could also compare the burden of *de novo* uORF-perturbing variants in cases with background rates predicted by a *de novo* mutational model. However, the challenge would be to take account of high methylation seen in 5'UTR, which is different from the sequence-context *de novo* mutational model for coding bases we have used in Chapter 3. With the growth of whole-genome sequencing data from UKBB and gnomAD, case-control burden testing might also be feasible to identify candidate genes with uORF-perturbing variants causing rare diseases.

5.5 Acknowledgements

- James Ware: study design and overall mentorship
- Nicola Whiffin; summarised the list of translated uORFs originally curated by the online repository of small ORFs (sorfs.org); analysed the candidate high-impact ClinVar 5'UTR VUS and generated Figure 5.2; study design and overall mentorship
- Genomics England 100K Genomes Project. This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support.

Chapter 6 Concluding Remarks

Overall, the thesis aimed to develop novel computational methods and tools to interpret genetic variants of rare diseases. To this end, I built a disease-specific variant classifier by incorporating genotype-phenotype relationship for inherited cardiac conditions; measured homologous residues intolerant of genetic variations; evaluated constrained *TTN* missense variants; and developed a variant annotator for 5’UTR variants.

6.1 What this work adds to the field

The development of variant pathogenicity prediction tools should be tailored to the needs of accurate genetic diagnosis

With a sufficient number of variants of known pathogenicity, we can use machine learning to generalise disease-causing patterns from computational lines of evidence such as using conservation, genetic constraint, and structural effect of residues for missense variants. The learned pattern could be applied to predict the pathogenicity of a novel variant. Though this general idea has been implemented in existing tools, In Chapter 2, I showed that the state-of-the-art genome-wide tools are imprecise to apply in variant interpretation: no consideration of gene-disease relationships and no standardized benchmarking relevant to clinical decision making.

Using inherited cardiac conditions as an example, a disease-specific model like CardioBoost could improve global classification accuracy by 4-24% over existing tools. Incorporating classification criteria with confidence level aligned with recommendations used for clinical practice, it could effectively prioritise variants of clinical relevance. CardioBoost provides a quantitative measure for supporting evidence PP3 and BP4 in ACMG guidelines, which could be applied in research and clinical laboratories to genetically diagnose inherited cardiac conditions.

This study also emphasizes that the model design in variant pathogenicity tools should be able to reflect real relationships in data and needs in genetic diagnosis. As intensively discussed, the model specificity might not be practical with limited training data. Given the size of interpreted variants for inherited cardiac conditions, we found a disease-specific model offered the optimal balance between model specificity and data availability compared with a gene-disease specific model or a genome-wide model. For other diseases, similar strategies could be experimented though it would require substantial efforts and knowledge on machine learning.

To democratise the usage of machine learning in variant pathogenicity prediction, future algorithmic developments could aim for an automated framework such that it can find an optimal model for a given training variant set of diseases. From an engineering perspective, this automated framework shall also be composable to satisfy the needs for different disease-specific models. It could also deploy transfer learning methods such that a prediction model developed previously for related tasks could also be reused in a new task. For example, in the context of variant pathogenicity prediction, deep learning models such as PrimateAI trained to learn conservation patterns across the genome could be deployed in a novel disease-specific prediction task as it has learned how to recognise conservation from sequences from large-scale training data. Then the deep neural network could be fine-tuned to recognise the specific pattern from a disease-specific training set even with a small size.

Homologous Residue Constraint is a novel computational line of evidence for missense variant interpretation.

To interpret variants of uncertain significance, I also seek to develop novel evidence to capture characteristics of variant pathogenicity that could be neglected by existing approaches. Motivated by the challenge of interpreting *TTN* missense variants, I developed HRC, a novel measure of genetic constraint at the level of single amino acids, to predict missense variants under purifying selection but sparsely distributed in genes. In benchmarking with existing genetic constraint scores, HRC has higher precision in predicting disease-causing variants. Applied in patients with developmental disorders, it can also discriminate pathogenic *de novo* missense variants from benign ones in disease genes with an effect size comparable to protein-truncating variants. With a novel line of evidence to prioritise more missense variants, the power of gene discovery could also be improved as shown in the examples of developmental disorders and hypertrophy cardiomyopathy. Compared with existing approaches, its statistical power does not depend on gene length or mutation clusters within a gene or sub-genic region but depends on the occurrence of a domain in the entire component of human proteins. Therefore, it provides an orthogonal measure of variant pathogenicity.

HRC can be immediately applied as a computational line of evidence to prioritise missense variants in clinics or research. In the assessment of novel variants, the constraint score could be combined with other computational tools. It would be our interest as future work to explore the best usage of HRC combining with other existing tools. It could also be integrated as a novel feature in the development of machine learning-based pathogenicity prediction tools. Since it is a domain-centric score, currently it is only applicable in proteins with Pfam domains.

This study also pinpoints measuring intraspecific purifying selection as a powerful resource to evaluate variant pathogenicity. Future curation of naturally occurred human genetic variation especially with increasing sample size and diversity of ancestries would be crucial to driving the accuracy and resolution of measurements.

A subset of *TTN* missense variants can be associated with DCM.

In previous studies, there wasn't any detectable association between *TTN* missense variants and DCM with population-level evidence. Rare *TTN* missense variants are collectively prevalent in the general population. Nearly every human being carries one. Though two variants have been identified through family linkage studies, there wasn't a way to identify a subset of disease-causing *TTN* missense variants. In Chapter 4, I applied the HRC method to prioritise DCM-associated *TTN* missense variants. The candidate positions are enriched with ClinVar pathogenic variants. The two DCM-causing variants known so far are located in the top constrained positions. *TTN* missense variants in I-set domains are also found to be significantly associated with DCM in Caucasian samples, mostly driven by cases of ACM. While we would need to replicate the association in independent datasets, the study represents a promising step forward and creates future research agenda.

This work also emphasizes that there can be a subset of *TTN* missense variants associated with DCM. For novel *TTN* missense variants found in patients of DCM, they should be recorded as well as patients' phenotype and familial data. As our analysis suggests that the candidate variants might explain 2% of DCM patients, data sharing across multiple cardiac centres would be strategic to aggregate samples to confirm the association robustly.

The interpretation of non-coding variant would benefit from improving the resolution of variant annotation

In Chapter 5, I presented UTRannotator, which could identify high-impact 5'UTR variants creating or disrupting uORFs. Out of all 5'UTR variants in ClinVar, we found that 32% of (likely) pathogenic variants and 8% of VUS can act through a uORF-disturbing mechanism.

We also found *de novo* uORF-disturbing variants in undiagnosed patients from GEL have a higher burden in LoF-intolerant genes compared with variants in gnomAD.

For genetic diagnosis, this plugin can be immediately included in the bioinformatic pipeline for variant annotation. Although whole-genome sequencing would be ideal, variants called from exome sequencing can also be analysed since exons in 5’UTR proximal to CDS are often detected in exome sequencing[152]. It has already been applied to identify uORF-perturbing variants on the exome sequencing data of 9,858 parent-offspring trios from the Deciphering Developmental Disorders study[152]. It is also extendable once we know how to predict the effects of other classes of UTR variants.

What are the lessons we learn from this work to scale up to other non-coding variant annotations?

One of the key factors to consider is the quality of functional non-coding region mapping. The annotation of UTR is a part of gene annotation, which is based on gene evidence (e.g. RNA-seq/Expressed sequence tags data) and protein evidence (e.g. ortholog in other species; experimental evidence of proteins). Initiatives such as MANE select (curating high-quality annotation by matching Ensembl and RefSeq transcripts) would largely harmonise the annotation of UTR. However, the annotation for the majority of non-coding regions still entails uncertainty, which is also complicated by tissue/cellular specificity.

The second factor is being able to predict the effects of non-coding genetic variations on disease. Our ability to directly predict the disease impact of non-coding variants vastly is limited by the fact that we only know the pathogenicity for a small fraction of non-coding variants (e.g. only 97 pathogenic/likely pathogenic 5’UTR variants in ClinVar), though it has been explored previously on certain non-coding regions[32,42,90,153]. Alternatively, we could focus on predicting the molecular effects of non-coding variants on regulatory elements. For

5'UTR variants, we show that a subset of them with putative loss-of-function effects on uORFs could be prioritised based on sequence changes. However, it is not straightforward to annotate variants disrupting sequence motifs in other cis-regulatory elements such as enhancers and promoters. The latter has been approached by pattern recognition from large-scale chromatin profiling data such as shown in DeepSEA[154]. High-throughput functional assays are also particularly important to elucidate the molecular effects especially for non-coding single nucleotide variations.

Finally, the study of high-impact 5'UTR variants wouldn't be possible without the ongoing curation of large-scale WGS datasets such as on undiagnosed patients by GEL and general populations by gnomAD. In the near future, we would also get better at establishing the link between large-effect non-coding variants and disease phenotypes by applying "genome-first" approaches with the increasing samples of WGS and clinical data in biobanks.

6.2 Conclusions

The work in the thesis has refined the accuracy of interpreting genetic variants causing inherited cardiac conditions, established an amino-acid level constraint to prioritise missense variants of medical relevance, identified promising *TTN* missense variants with a role in DCM, and enhanced the identification of molecular consequence of 5'UTR variants. In summary, these studies have generated widely applicable methods and tools to study genetic diseases and improved understanding of inherited cardiac conditions.

References

1. Belkadi, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 5473–5478 (2015).
2. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405–423 (2015).
3. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6**, R44 (2005).
4. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, (2016).
5. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* **6**, 80–92 (2012).
6. Geraldine A. Van der Auwera, B. D. O. *Genomics in the Cloud [Book]*. (O'Reilly Media, Inc., 2020).
7. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science (80-.).* **335**, 823–828 (2012).
8. Walsh, R. *et al.* Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet Med* **19**, 192–203 (2017).
9. Pagani, F., Raponi, M. & Baralle, F. E. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 6368–6372 (2005).
10. Plotkin, J. B. & Kudla, G. Synonymous but not the same: The causes and consequences of codon bias. *Nature Reviews Genetics* **12**, 32–42 (2011).

11. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1082 (2009).

12. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, (2010).

13. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).

14. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Publ. Gr.* **12**, (2011).

15. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50**, 1161–1170 (2018).

16. Frazer, J. *et al.* Large-scale clinical interpretation of genetic variants using evolutionary data and deep learning. *bioRxiv* 2020.12.21.423785 (2020). doi:10.1101/2020.12.21.423785

17. Lal, D. *et al.* Gene family information facilitates variant interpretation and identification of disease-associated genes in neurodevelopmental disorders. *Genome Med.* (2020). doi:10.1186/s13073-020-00725-6

18. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).

19. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

20. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* (2020). doi:10.1038/s41586-020-2308-7

21. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

22. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics* **12**, 745–755 (2011).

23. Whiffin, N. *et al.* Using high-resolution variant frequencies to empower clinical

genome interpretation. *Genet. Med.* **19**, 1151–1158 (2017).

24. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* (2014). doi:10.1038/ng.3050

25. Petrovski, S. *et al.* The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. *PLoS Genet.* **11**, (2015).

26. Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* (2017). doi:10.1101/148353

27. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* (2019). doi:10.1038/s41588-018-0294-6

28. Gussow, A. B., Petrovski, S., Wang, Q., Allen, A. S. & Goldstein, D. B. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol.* **17**, (2016).

29. Traynelis, J. *et al.* Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation. *Genome Res.* **27**, 1715–1729 (2017).

30. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276–283 (2015).

31. Whiffin, N. *et al.* Characterising the loss-of-function impact of 5’ untranslated region variants in 15,708 individuals. *Nat. Commun.* (2020). doi:10.1038/s41467-019-10717-9

32. Vitsios, D., Dhindsa, R. S., Middleton, L., Gussow, A. B. & Petrovski, S. Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. *Nat. Commun.* **12**, 1–14 (2021).

33. Ittisoponpisan, S. *et al.* Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *J. Mol. Biol.* **431**, 2197–2212 (2019).

34. Ponzoni, L. & Bahar, I. Structural dynamics is a determinant of the functional significance of missense variants. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4164–4169 (2018).

35. Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R. & Alexov, E. Molecular mechanisms of disease-causing missense mutations. *Journal of Molecular Biology* **425**, 3919–3936 (2013).

36. Homburger, J. R. *et al.* Multidimensional structure-function relationships in human β-cardiac myosin from population-scale genetic variation. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 6701–6706 (2016).

37. Stephenson, J. D., Laskowski, R. A., Nightingale, A., Hurles, M. E. & Thornton, J. M. VarMap: a web tool for mapping genomic coordinates to protein sequence and structure and retrieving protein structural annotations. *Bioinformatics* **35**, 4854–4856 (2019).

38. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).

39. Jagadeesh, K. A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* **48**, 1581–1586 (2016).

40. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).

41. Qi, H. *et al.* MVP predicts the pathogenicity of missense variants by deep learning. *Nat. Commun.* (2021). doi:10.1038/s41467-020-20847-0

42. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* **48**, 214–220 (2016).

43. Kishore Jaganathan, A. *et al.* Predicting Splicing from Primary Sequence with Deep Learning In Brief A deep neural network precisely models mRNA splicing from a genomic sequence and accurately predicts noncoding cryptic splice mutations in patients with rare genetic diseases. Predicting Splicing from Primary Sequence with

Deep Learning. *Cell* **176**, 535–548 (2018).

44. McKenna, W. J., Maron, B. J. & Thiene, G. Classification, epidemiology, and global burden of cardiomyopathies. *Circ. Res.* **121**, 722–730 (2017).

45. Rosenbaum, A. N., Agre, K. E. & Pereira, N. L. Genetics of dilated cardiomyopathy: practical implications for heart failure management. *Nature Reviews Cardiology* **17**, 286–297 (2020).

46. Hershberger, R. E., Hedges, D. J. & Morales, A. Dilated cardiomyopathy: The complexity of a diverse genetic architecture. *Nature Reviews Cardiology* **10**, 531–547 (2013).

47. Ware, J. S. & Cook, S. A. Role of titin in cardiomyopathy: From DNA variants to patient stratification. *Nature Reviews Cardiology* (2018). doi:10.1038/nrcardio.2017.190

48. McNally, E. M. & Mestroni, L. Dilated cardiomyopathy: Genetic determinants and mechanisms. *Circulation Research* **121**, 731–748 (2017).

49. Fatkin, D. *et al.* Titin truncating mutations: A rare cause of dilated cardiomyopathy in the young. *Prog. Pediatr. Cardiol.* **40**, 41–45 (2016).

50. Mazzarotto, F. *et al.* Reevaluating the Genetic Contribution of Monogenic Dilated Cardiomyopathy. *Circulation* 387–398 (2020). doi:10.1161/CIRCULATIONAHA.119.037661

51. Rehm, H. L. *et al.* ClinGen — The Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).

52. Ware, J. S. *et al.* Shared Genetic Predisposition in Peripartum and Dilated Cardiomyopathies. *N. Engl. J. Med.* **374**, 233–241 (2016).

53. Marian, A. J. & Braunwald, E. Hypertrophic cardiomyopathy: Genetics, pathogenesis, clinical manifestations, diagnosis, and therapy. *Circ. Res.* **121**, 749–770 (2017).

54. Maron, B. J. *et al.* Prevalence of hypertrophic cardiomyopathy in a general population of young adults: Echocardiographic analysis of 4111 subjects in the CARDIA study. *Circulation* **92**, 785–789 (1995).

55. Maron, B. J., Mathenge, R., Casey, S. A., Poliac, L. C. & Longe, T. F. Clinical profile of hypertrophic cardiomyopathy identified de novo in rural communities. *J. Am. Coll. Cardiol.* **33**, 1590–1595 (1999).

56. Ingles, J. *et al.* Evaluating the Clinical Validity of Hypertrophic Cardiomyopathy Genes. *Circ. Genomic Precis. Med.* **12**, (2019).

57. Ho, C. Y. *et al.* Genotype and lifetime burden of disease in hypertrophic cardiomyopathy: insights from the Sarcomeric Human Cardiomyopathy Registry (SHaRe). *Circulation* **138**, 1387–1398 (2018).

58. de Marvao MRCP, A. *et al.* Outcomes and phenotypic expression of rare variants in hypertrophic cardiomyopathy genes amongst UK Biobank participants □-joint senior / corresponding authors. *medRxiv* 2021.01.21.21249470 (2021). doi:10.1101/2021.01.21.21249470

59. Tadros, R. *et al.* Shared genetic pathways contribute to risk of hypertrophic and dilated cardiomyopathies with opposite directions of effect. *Nat. Genet.* **53**, 128–134 (2021).

60. Pirruccello, J. P. *et al.* Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. *Nat. Commun.* **11**, 1–10 (2020).

61. Harper, A. R. *et al.* Common genetic variants and modifiable risk factors underpin hypertrophic cardiomyopathy susceptibility and expressivity. *Nat. Genet.* **53**, 135–142 (2021).

62. Schwartz, P. J. *et al.* Prevalence of the congenital long-qt syndrome. *Circulation* **120**, 1761–1767 (2009).

63. Sarquella-Brugada, G., Campuzano, O., Arbelo, E., Brugada, J. & Brugada, R. Brugada syndrome: Clinical and genetic findings. *Genetics in Medicine* **18**, 3–12 (2016).

64. Adler, A. *et al.* An International, Multicentered, Evidence-Based Reappraisal of Genes Reported to Cause Congenital Long QT Syndrome. *Circulation* **141**, 418–428 (2020).

65. Kapa, S. *et al.* Genetic testing for long-QT syndrome: distinguishing pathogenic mutations from benign variants. *Circulation* **120**, 1752–1760 (2009).

66. Hosseini, S. M. *et al.* Reappraisal of reported genes for sudden arrhythmic death: Evidence-based evaluation of gene validity for brugada syndrome. *Circulation* **138**, 1195–1205 (2018).

67. Kapplinger, J. D. *et al.* An international compendium of mutations in the SCN5A-encoded cardiac sodium channel in patients referred for Brugada syndrome genetic testing. *Hear. Rhythm* **7**, 33–46 (2010).

68. Anderson, D., Baynam, G., Blackwell, J. M. & Lassmann, T. Personalised analytics for rare disease diagnostics. *Nat. Commun.* (2019). doi:10.1038/s41467-019-13345-5

69. Ruklisa, D., Ware, J. S., Walsh, R., Balding, D. J. & Cook, S. A. Bayesian models for syndrome- and gene-specific probabilities of novel variant pathogenicity. *Genome Med.* **7**, 5 (2015).

70. Pugh, T. J. *et al.* The landscape of genetic variation in dilated cardiomyopathy as surveyed by clinical DNA sequencing. *Genet. Med.* **16**, 601–608 (2014).

71. Alfares, A. A. *et al.* Results of clinical genetic testing of 2,912 probands with hypertrophic cardiomyopathy: Expanded panels offer limited additional sensitivity. *Genet. Med.* **17**, 880–888 (2015).

72. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).

73. Aguib, Y. *et al.* The Egyptian Collaborative Cardiac Genomics (ECCO-GEN) Project: defining a healthy volunteer cohort. *npj Genomic Med.* (2020). doi:10.1038/s41525-020-00153-w

74. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

75. Weile, J. *et al.* A framework for exhaustively mapping functional missense variants. *Mol. Syst. Biol.* (2017). doi:10.15252/msb.20177908

76. Zhang, J. *et al.* Assessing predictions on fitness effects of missense variants in

calmodulin. *Hum. Mutat.* (2019). doi:10.1002/humu.23857

77. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).

78. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**, 10915–10919 (1992).

79. Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science.* **185**, 862–864 (1974).

80. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).

81. Siepel, A., Pollard, K. S. & Haussler, D. *New Methods for Detecting Lineage-Specific Selection*. *Research in Computational Molecular Biology* (Springer Berlin Heidelberg, 2006).

82. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).

83. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25**, (2009).

84. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **76**, 7.20.1-7.20.41 (2013).

85. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. Mutationtaster2: Mutation prediction for the deep-sequencing age. *Nature Methods* **11**, 361–362 (2014).

86. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Res.* **39**, (2011).

87. Shihab, H. A. *et al.* Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum. Genomics* **8**, (2014).

88. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* **7**, (2012).

89. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. *BMC Genomics* **14**, S3 (2013).

90. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

91. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761 (2014).

92. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24**, 2125–2137 (2015).

93. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).

94. Schafer, J. L. & Graham, J. W. Missing data: Our view of the state of the art. *Psychol. Methods* **7**, 147–177 (2002).

95. R Core Team. R: A Language and Environment for Statistical Computing. (2017).

96. Bischl, B. *et al.* mlr: Machine Learning in R. *J. Mach. Learn. Res.* **17**, 1–5 (2016).

97. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification and Regression Trees*. (Taylor & Francis, 1984).

98. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**, 21–27 (1967).

99. Zou, H. & Hastie, T. Regularization and variable selection via the elastic-net. *J. R. Stat. Soc.* **67**, 301–320 (2005).

100. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning: data mining, inference and prediction*. (Springer, 2009).

101. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).

102. Chipman, H. A., George, E. I. & McCulloch, R. E. BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **6**, 266–298 (2012).

103. Chen, T. & Guestrin, C. *XGBoost : Reliable Large-scale Tree Boosting System*.

Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM, 2016).

104. Cawley, G. C. & Talbot, N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.* (2010).

105. Boughorbel, S., Jarray, F. & El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One* **12**, (2017).

106. Cirino, A. L. *et al.* Role of genetic testing in inherited cardiovascular disease: A review. *JAMA Cardiol.* **2**, 1153–1160 (2017).

107. Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Min.* **10**, (2017).

108. Good, P. I. *Resampling methods : a practical guide to data analysis*. (Birkhäuser, 2010).

109. Kapplinger, J. D. *et al.* Spectrum and prevalence of mutations from the first 2,500 consecutive unrelated patients referred for the FAMILION long QT syndrome genetic test. *Hear. Rhythm* **6**, 1297–1303 (2009).

110. Kapplinger, J. D. *et al.* An international compendium of mutations in the SCN5A-encoded cardiac sodium channel in patients referred for Brugada syndrome genetic testing. *Hear. Rhythm* **7**, 33–46 (2010).

111. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, (2015).

112. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **44**, 837–845 (1988).

113. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD®): 2003 Update. *Hum. Mutat.* **21**, 577–581 (2003).

114. Lopes, L. R., Rahman, M. S. & Elliott, P. M. A systematic review and meta-analysis of genotype-phenotype associations in patients with hypertrophic cardiomyopathy

caused by sarcomeric protein mutations. *Heart* **99**, 1800–1811 (2013).

115. Ingles, J. *et al.* Nonfamilial Hypertrophic Cardiomyopathy. *Circ. Cardiovasc. Genet.* **10**, (2017).

116. Al-Numair, N. S. *et al.* The structural effects of mutations can aid in differential phenotype prediction of beta-myosin heavy chain (Myosin-7) missense variants. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw362

117. Jordan, D. M. *et al.* Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy. *Am. J. Hum. Genet.* (2011). doi:10.1016/j.ajhg.2011.01.011

118. Evans, P. *et al.* Genetic variant pathogenicity prediction trained using disease-specific clinical sequencing data sets. *Genome Res.* (2019). doi:10.1101/gr.240994.118

119. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genet.* **9**, (2013).

120. Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* (2018). doi:10.1038/nature25983

121. Strumillo, M. J. *et al.* Conserved phosphorylation hotspots in eukaryotic protein domain families. *Nat. Commun.* (2019). doi:10.1038/s41467-019-09952-x

122. Mistry, J., Bateman, A. & Finn, R. D. Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics* (2007). doi:10.1186/1471-2105-8-298

123. Wiel, L., Venselaar, H., Veltman, J. A., Vriend, G. & Gilissen, C. Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics. *Hum. Mutat.* (2017). doi:10.1002/humu.23313

124. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky995

125. Satterstrom, F. K. *et al.* Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell* (2020). doi:10.1016/j.cell.2019.12.036

126. Kaplanis, J. *et al.* Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* (2020). doi:10.1038/s41586-020-2832-5

127. Campbell, C. D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* **44**, (2012).

128. Kosmicki, J. A. *et al.* Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* (2017). doi:10.1038/ng.3789

129. Ware, J. S., Samocha, K. E., Homsy, J. & Daly, M. J. Interpreting de novo Variation in Human Disease Using denovolyzeR. *Curr. Protoc. Hum. Genet.* (2015). doi:10.1002/0471142905.hg0725s87

130. Thormann, A. *et al.* Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat. Commun.* **10**, 2373 (2019).

131. Esposito, D. *et al.* MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol.* **20**, 223 (2019).

132. Schafer, S. *et al.* Titin-truncating variants affect heart function in disease cohorts and the general population. *Nat. Genet.* (2017). doi:10.1038/ng.3719

133. Gerull, B. *et al.* Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. *Nat. Genet.* (2002). doi:10.1038/ng815

134. Halaby, D. M., Poupon, A. & Mornon, J.-P. The immunoglobulin fold family: sequence analysis and 3D structure comparisons. *Protein Eng. Des. Sel.* **12**, 563–571 (1999).

135. Hinson, J. T. *et al.* Titin mutations in iPS cells define sarcomere insufficiency as a cause of dilated cardiomyopathy. *Science (80-.).* (2015). doi:10.1126/science.aaa5458

136. Akinrinade, O. *et al.* Relevance of Titin Missense and Non-Frameshifting Insertions/Deletions Variants in Dilated Cardiomyopathy. *Sci. Rep.* (2019). doi:10.1038/s41598-019-39911-x

137. Begay, R. L. *et al.* Role of titin missense variants in dilated cardiomyopathy. *J. Am. Heart Assoc.* (2015). doi:10.1161/JAHA.115.002645

138. Herrero Galán, E. *et al.* Conserved cysteines in titin sustain the mechanical function of cardiomyocytes. *bioRxiv* (2020).

139. Hedberg, C. *et al.* Hereditary myopathy with early respiratory failure is associated with misfolding of the titin fibronectin III 119 subdomain. *Neuromuscul. Disord.* (2014). doi:10.1016/j.nmd.2014.02.003

140. Steele, H. E. *et al.* Cardiac involvement in hereditary myopathy with early respiratory failure. *Neurology* (2016). doi:10.1212/WNL.0000000000003064

141. Laddach, A., Gautel, M. & Fraternali, F. TITINdb—a computational tool to assess titin's role as a disease gene. *Bioinformatics* **33**, 3482–3485 (2017).

142. Heyne, H. O. *et al.* Predicting functional effects of missense variants in voltage-gated sodium and calcium channels. *Sci. Transl. Med.* **12**, 6848 (2020).

143. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science* (2016). doi:10.1126/science.aad9868

144. Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. U. S. A.* (2009). doi:10.1073/pnas.0810916106

145. Kozak, M. The scanning model for translation: An update. *Journal of Cell Biology* (1989). doi:10.1083/jcb.108.2.229

146. Chen, J. *et al.* Pervasive functional translation of noncanonical human open reading frames. *Science (80-.).* (2020). doi:10.1126/science.aav5912

147. *The 100,000 Genomes Project Protocol v3, Genomics England.* (2017).

148. Olexiouk, V. *et al.* SORFs.org: A repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkv1175

149. McGillivray, P. *et al.* A comprehensive catalog of predicted functional upstream open reading frames in humans. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gky188

150. Lee, D. S. M. *et al.* Disrupting upstream translation in mRNAs is associated with human disease. *Nat. Commun.* **12**, 1–14 (2021).

151. Fischbach, G. D. & Lord, C. The simons simplex collection: A resource for identification of autism genetic risk factors. *Neuron* **68**, 192–195 (2010).

152. Wright, C. F. *et al.* Non-coding variants upstream of MEF2C cause severe developmental disorder through three distinct loss-of-function mechanisms. doi:10.1101/2020.11.15.20229807

153. Caron, B., Luo, Y. & Rausell, A. NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. *Genome Biol.* **20**, 32 (2019).

154. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).